

REPORT DOCUMENTATION PAGE				
1a. REPORT SECURITY CLASSIFICATION Unclassified			1b. RESTRICTIVE MARKINGS	
2a. SECURITY CLASSIFICATION AUTHORITY NA			3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.	
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE NA				
4. PERFORMING ORGANIZATION REPORT NUMBER(S) IDA Paper P-1922			5. MONITORING ORGANIZATION REPORT NUMBER(S)	
6a. NAME OF PERFORMING ORGANIZATION Institute for Defense Analyses		6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION DoD-IDA Management Office	
6c. ADDRESS (City, State, and Zip Code) 1801 N. Beauregard St. Alexandria, VA 22311			7b. ADDRESS (CITY, STATE, AND ZIP CODE) 1801 N. Beauregard St. Alexandria, VA 22311	
8a. NAME OF FUNDING/SPONSORING ORGANIZATION ASD (FMP)/TPD		8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA 903 79 C 0018 MDA 903 84 C 0031	
8c. ADDRESS (City, State, and Zip Code) The Pentagon Washington, DC 20301			10. SOURCE OF FUNDING NUMBERS	
			PROGRAM ELEMENT	PROJECT NO.
11. TITLE (Include Security Classification) Performance Measures for Evaluating the Effectiveness of Maintenance Training (U)				
12. PERSONAL AUTHOR(S) Richard S. Gibson, Jesse Orlansky				
13. TYPE OF REPORT Final		13b. TIME COVERED FROM 9/85 TO 2/86	14. DATE OF REPORT (Year, Month, Day) September 1986	
15. PAGE COUNT 75				
16. SUPPLEMENTARY NOTATION				
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number) Training-effectiveness, maintenance training, simulators, simulated aviation maintenance training, training research, performance measurement	
FIELD	GROUP	SUB-GROUP		
19. ABSTRACT (Continue on reverse if necessary and identify by block number) This paper reviews and summarizes 17 studies concerned with the collection of objective performance data for the purpose of evaluating the effectiveness of maintenance training. The results are discussed in relation to five major aspects of evaluating the effectiveness of training: 1) transfer of training, 2) quality of simulator, 3) effects of training on individual performance, 4) differential effects of alternative methods of training, and 5) the effects of training and experience on unit performance and operational readiness. Based on the analyses, a number of areas for future research and development were suggested.				
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT <input type="checkbox"/> UNCLASSIFIED/UNLIMITED <input checked="" type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED	
22a. NAME OF RESPONSIBLE INDIVIDUAL Jesse Orlansky			22b. TELEPHONE (Include Area Code) (703) 845-2293	
			22c. OFFICE SYMBOL	

Gibson

Copy

121 of 165 copies

LIBRARY
RESEARCH REPORTS DIVISION
NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA 93940

IDA PAPER P-1922

PERFORMANCE MEASURES FOR EVALUATING THE EFFECTIVENESS OF MAINTENANCE TRAINING

✓
Richard S. Gibson
Jesse Oriansky

September 1986

Prepared for
Office of the Assistant Secretary of Defense
(Force Management and Personnel)



INSTITUTE FOR DEFENSE ANALYSES
1801 N. Beauregard Street, Alexandria, VA 22311

IDA PAPER P-1922

PERFORMANCE MEASURES FOR EVALUATING THE
EFFECTIVENESS OF MAINTENANCE TRAINING

Richard S. Gibson
Jesse Orlansky

September 1986



INSTITUTE FOR DEFENSE ANALYSES

Contract MDA 903 79 C 0018
Task T-3-167

Contract MDA 903 84 C 0031
Task T-L2-308

FOREWORD

This report was prepared by the Institute for Defense Analyses (IDA) under Contract MDA 903 79 C 0018, Task Order T-3-167, and Contract MDA 903 84 C 0031, Task Order T-L2-308. The cognizant technical officer for these tasks is Mr. Gary Boycan, Assistant Director, Training Systems and Technology, Office of the Assistant Secretary of Defense (Force Management and Personnel)/Training Policy Directorate.

The training effectiveness measures reviewed in the preparation of this report were limited to the area of maintenance training. However, many of the approaches, measurement techniques, and technical insights are applicable to the evaluation of a broad range of training effectiveness issues. Most of the literature, which has focused on the early effects of training, has only marginal usefulness in formulating training policy. Increased effort should be directed toward assessing the long-term effects of training and experience on the quality of individual performance, unit effectiveness, and ultimately, on combat readiness.

ACKNOWLEDGMENTS

The many contributions to this paper by colleagues and friends who provided technical advice and detailed comments are gratefully acknowledged. We are indebted to all those named below who helped improve this paper, and hope they will overlook any failures to follow their advice. The authors alone are responsible for what appears in this paper.

Earl A. Alluisi	Air Force Human Resources Laboratory
Clyde A. Bricton	CB Associates
Kenneth D. Cross	Anacapa Sciences
Lyle M. Ehrenberg	Center for Competency Development
Gary Green	Defense Training Data and Analysis Center
R. Bruce Gould	Air Force Human Resources Laboratory
Jack H. Hiller	ARI Field Unit-Monterey
Norman E. Lane	Essex Corporation
Rodney D. McConnell	Management Consulting and Research, Inc.
Ruth A. Wienclaw	Wienclaw Associates
Emily Willey	Honeywell, Inc.

We appreciate the assistance provided by Phyllis M. Hamrick, Anthea DeV Vaughan, and Brenda A. Jaggie in the preparation of this paper.

CONTENTS

FOREWORD.....	iii
ACKNOWLEDGMENTS.....	v
LIST OF ABBREVIATIONS.....	xiii
SUMMARY.....	S-1
CONCLUSIONS.....	C-1
I. INTRODUCTION.....	1
A. Observe and Document Actual Job Performance.....	4
B. Observe Performance on Job Samples Under Controlled Conditions.....	4
C. Assess the Effects of Training From Information in Existing Data Banks.....	5
II. DATA ON JOB PERFORMANCE	7
A. Job Performance Data Collected Under Controlled Field Conditions.....	7
B. Job Performance Data Collected Under Field Survey Conditions.....	15
C. Use of Simulators to Collect Job Performance Data.....	20
D. Data Relevant to Training Derived from Existing Data Banks.....	26
III. DISCUSSION	37
A. Sources of Training Effectiveness Measurement and Evaluation.....	37
B. Integration of Multiple Measures of F-16 Maintenance Training Effectiveness to Better Understand the Measures and the Systems.....	49
C. Job Performance Measures for Evaluating Training Effectiveness.....	54
REFERENCES	61

FIGURES

Figure 1.	Troubleshooting Performance as a Function of Experience.....	10
Figure 2.	Performance Ratings (BARS Scores) for SAMT- and AET-trained Students and Technicians.....	13
Figure 3.	Productivity Growth for Electricians' Mates in the First-Term Enlistment.....	19
Figure 4.	Mean Training Times Needed by Groups Trained in a Tank or Simulator to Reach Criterion Performance.....	21
Figure 5.	Student Confidence in their Ability to Perform Tasks on the Simulator at End of Course.....	25
Figure 6.	Proportion of Errors on Tasks Performed on Simulator at End of Course.....	25
Figure 7.	Effects of Training on Turbofan Power Plant (WUC 23Z00) Maintenance.....	30
Figure 8.	Effects of Training on Instrument and Flight Control Systems (WUC 14A00) Maintenance.....	30
Figure 9.	Aggregate Training Effects on Turbofan Power Plat (WUC 23000) Maintenance Productivity.....	32
Figure 10.	Aggregate Training Effects on Electrical Power Supply (WUC 42000) Maintenance Productivity.....	32
Figure 11.	Aggregate Training Effects on Flight Control Systems (WUC 14000) Maintenance Productivity.....	33
Figure 12.	Effects of FTD Training Methods on Time Needed to Complete Specific Maintenance Tasks.....	34

TABLES

Table 1.	Classification Scheme for Categorizing Job Performance Measures Used to Assess Training Effectiveness.....	3
Table 2.	Troubleshooting Interview Ratings for TD- and AET-Trained Technicians.....	9
Table 3.	Correlations Between Performance Ratings (BARS) of Course Graduates as Students and Technicians.....	11
Table 4.	Statistical Evaluation of Improvements in BARS Scores Over Time (F Values for Repeated Measures ANOVA).....	12
Table 5.	Intercorrelations Between Walk Through Performance Test (WTPT) Scores and Component Subtest Scores.....	14
Table 6.	Multiple Regression of Performance and Training Variables on WTPT Performance.....	14
Table 7.	Quality Assurance Personnel Test Results.....	16
Table 8.	Task Completion Time Ratios.....	17
Table 9.	Mean Values of Experienced Tank Drivers on Four Tasks.....	21
Table 10.	F-16 Simulators and Training Devices.....	23
Table 11.	Summary of Instructor Ratings of Simulator Fidelity and Student Confidence Ratings and Performance Errors.....	24
Table 12.	Summary of Instructor Ratings of the Comparative Fidelity of Operational Checks and Fault Isolation Checks and Student Confidence Ratings and Performance Errors.....	24
Table 13.	Marginal Product of Paygrade Groups in Generating A-7 Sorties Quarter.....	28
Table 14.	Summary of Job Performance Measures of Maintenance.....	38
Table 15.	Measures of F-16 Maintenance Training Effectiveness Summarized by Simulator System.....	50
Table 16.	Summary of Job Performance Measures and Studies Used to Assess the Training Effectiveness of F-16 Maintenance Training Devices.....	55
Table 17.	Job Performance Measurement Characteristics.....	57

LIST OF ABBREVIATIONS

AET	Actual Equipment Trainer
AFB	Air Force Base
AFSC	Air Force Specialty Code
ANOVA	Analysis of variance
BARS	Behaviorally Anchored Rating Scale
CASREP	Casualty Report
CCD	Center for Competency Development
CRTRNG	Combat readiness training
DMR	Desired Maintenance Results
FTD	Field Training Detachment
HT	Hardware trainer
JPD	Job performance measured directly
JPI	Job performance measured indirectly
JPMS	Job Performance Measurement System
MDC	Maintenance Data Collection
MMICS	Maintenance Management Information and Control System
O	Objective
OJT	On-the-Job Training
PD	Personnel performance measured directly
PI	Personnel performance measured indirectly
QA	Quality Assurance Personnel Test
RETOK	Retest OK
S	Subjective
SAMT	Simulated Aircraft Maintenance Trainer
SME	Subject matter expert
TD	Training device
UNITREP	Unit Status and Identity Report
USAF	United States Air Force
WTPT	Walk Through Performance Test
WUC	Work Unit Code

SUMMARY

A. PURPOSE

The intent of this paper is to review and summarize the current literature reporting the use of operational job performance measures to evaluate the effectiveness of maintenance training. The major results of the review are:

- (1) The identification and description of the kinds of job performance measures available, including a classification structure to increase the ease of ordering and understanding the various measures presented in the literature.
- (2) An analysis of the kinds of training effectiveness information that may be obtained through the use of specific job performance measures.
- (3) The presentation of some directions to be pursued in order to regularly and routinely evaluate the effectiveness of maintenance training programs.

An overview of the relevant research and a knowledge of the results of using performance data to evaluate training effectiveness are considered as necessary precursors to the development of an adequate training effectiveness methodology.

B. BACKGROUND

The purpose of military training is to prepare people to perform the technical tasks necessary to assure the availability and proper functioning of military weapons systems and support equipment. Without credible information about how well the graduates of training courses perform after leaving school, it is not possible to determine how well the courses provide the knowledge and skills needed to perform on the job. Until now, the effectiveness of training has been evaluated primarily by end-of-course tests or job-sample tests, which are indirect measures, or by supervisors' ratings of job performance, which are subjective measures. More recently, a number of research efforts have extended the range of available measures to include objective measures of job performance and better

controlled rating scales. It is now possible to provide a greater understanding of the advantages and limitations of a number of these different types of job performance measures for use in evaluating maintenance training effectiveness.

C. SCOPE

This paper contains a review and analysis of recent efforts to collect objective job performance data for the purpose of evaluating the effectiveness of maintenance training. The measurement techniques and results of 17 studies conducted since 1977 have been analyzed; most of the data reported here have been published since 1983. To assist in the analysis process, the reported training effectiveness data were sorted into categories in terms of whether the measures of job performance were subjective, observed directly, or inferred indirectly by analyzing relevant available data and determining whether individual or group performance data were used. The results are discussed in relation to five major aspects of evaluating the effectiveness of training:

- Transfer of training,
- Quality of simulation,
- Effects of training on individual performance,
- Differential effects of alternative methods of training,
- The effects of training and experience on unit performance and operational readiness.

CONCLUSIONS

1. OBJECTIVE MEASURES

Maintenance management Work Unit Code (WUC) data collected at the Work Center can provide objective information on speed of work that is of great value in evaluating the effects of training and training methods on actual performance of maintenance in the field.

2. SUBJECTIVE MEASURES

When objective job performance data are not reasonably obtainable, subjective measures, such as behaviorally anchored rating scales (BARS) and net productivity estimates, yield useful information on the effect of training on job performance.

3. EXPERIENCE READINESS

Training establishes an initial level of proficiency and provides a base for additional learning. The data show that differences in training background influence the rate at which technicians improve with on-the-job experience: this phenomenon has been termed the "experience readiness" effect. Where manifested, this effect is highly significant. Therefore, estimates of training effectiveness should include not only measures of initial maintenance performance but also measures of the rate at which proficiency increases as a function of on-the-job experience.

4. SIMULATED AVIATION MAINTENANCE TRAINERS (SAMTs)

SAMTs appear to be as effective as actual equipment trainers (AETs) for training maintenance technicians, as measured by their on-the-job performance in the field. The observed level of effectiveness varies from simulator to simulator and from task to task within a simulator. Based on objective data, some of the SAMT simulators were found to

be more effective than the AETs, some were about equally effective, and some were less effective. The effectiveness of the SAMTs was closely related to instructor ratings of simulator fidelity; i.e., those simulators which had the highest fidelity ratings also seemed to be the most effective. SAMT-trained personnel were consistently better at performing the Test-Inspect-Service tasks, while the AET-trained personnel were consistently better at performing the Remove-and-Replace tasks.

5. MAINTENANCE TRAINING EFFECTIVENESS MEASURES

No single maintenance performance measure can fulfill all requirements for evaluating the effectiveness of training. Measures must be selected on the basis of their availability, subjectivity, or objectivity, and whether they directly or indirectly measure individual or job performance. Direct objective measures of job performance, such as those that can be obtained from maintenance management data, have a significant potential for providing improved information concerning the specific benefits and weaknesses of alternative methods of training.

6. IMPORTANCE OF MAINTENANCE TRAINING

Data from Navy Casualty Reports (CASREPs) indicate that the formal training and experience-induced training of maintenance personnel are significant predictors of combat readiness.

7. IMPORTANCE OF SIMULATION QUALITY

Student confidence and performance closely parallel instructor ratings of simulator fidelity. This observation was reinforced by job performance data at the task level, which provided a profile of the strengths and weaknesses of the simulators evaluated. Interpretation of any training effectiveness evaluation of a simulated maintenance trainer depends in part on an understanding of the device's behavioral fidelity on critical tasks. To make any generalizations about the effectiveness of simulator-based training without considering the fidelity of the simulators would be unwarranted.

8. STATE OF THE ART OF MAINTENANCE PERFORMANCE MEASUREMENT FOR USE IN EVALUATING TRAINING EFFECTIVENESS

There are currently no proven off-the-shelf methodologies for collecting job-performance data to evaluate the effectiveness of maintenance training. However, several areas, listed below, deserve consideration for continued growth and development:

- More extensive use of improved rating scales such as the Behaviorally Anchored Rating Scales (BARS) or Net Job Productivity Ratings.
- Continued development of job sample tests such as the Walk Through Performance Test (WTPT). Of particular interest would be job sample tests which have been validated with objective data from maintenance management data banks.
- Continued effort to explore the possibility of using maintenance management data for evaluating training effectiveness appears to be justified by the value of the data when it is in usable form. The difficulty in obtaining usable data is an area that needs further exploration.
- Investigations using multiple performance-assessment techniques are needed to establish the comparability and relative effectiveness of the many methods currently being used. Application of a common set of measures would be more productive than the current practice of developing a new set of measures for every study.
- Continued studies to relate maintenance training to macro-level results such as unit performance or combat readiness are needed to provide better training management information.

I. INTRODUCTION

The effectiveness of training can be inferred from readily available school data such as students' grades on tests, percentage of students who pass a course, percentage of content mastery, and learning time. Although such measures are readily available, they are not the most relevant indicators of training effectiveness..

The purpose of military training is to prepare people to perform various jobs and not, in any general sense, to complete courses at school. Unless we have credible information about how well graduates of particular courses perform after leaving school, we do not know very much about whether the course provided the information needed to perform well on the job or whether, even if the course material was highly relevant, it was provided in such a way that success at school contributes to success on the job.

The conditions under which data are collected in the field and the types of data used to measure maintenance performance contribute to the kinds of inferences that can be drawn with regard to training effectiveness. The conditions under which the field data are collected can range from controlled experiments to field conditions that approximate an experiment to field surveys. Since different collection conditions vary with respect to the degree of experimental control exercised, there are corresponding differences in the credibility of the data, and the extent to which causal inferences can be drawn from the data.

Each of the studies reviewed used one or more unique methods for assessing the performance effectiveness of maintenance personnel. To make this heterogeneous collection of performance measures more manageable and understandable, the data were partitioned into a set of categories based upon a common group of features found within all of the data collection paradigms reviewed.

Each of the subject job-performance measures is the composite result of its functional relationship to three factors intrinsic to the measurement situation: (1) the person or persons doing the measuring, (2) the individual or individuals being measured, and (3) the task or job performance represented by the measure. Based upon this three-part

concept of performance measurement, all of the measures reviewed were categorized in accordance with the following guidelines. If the measure was heavily dependent upon individual interpretation and judgment, such as supervisor or peer ratings, it was classified as subjective (S). However, if the measure was largely independent of individual interpretation such as a speed- or accuracy-of-performance record or a test score, it was classified as objective (O).

If the measure was the result of the measurer's direct observation or experience with the individuals or personnel performing the task, such as a supervisor's rating, it was classified as a personnel direct (PD) measure. However, if the effectiveness of the individual or group performance was inferred from some result such as work hours to completion or comparative rates of A-7 flights off a carrier, it was classified as a personnel indirect (PI) measure.

If the measure used actually recorded the quantity or quality of the maintenance technician's job performance, such as a supervisor rating of job performance or work hours to completion, it was classified as a direct job performance (JPD) measure. However, if the effectiveness of job performance was inferred through the use of either a surrogate measure such as a job sample test or a job consequence such as the rate of A-7 sorties, it was classified as an indirect job performance (JPI) measure.

This three-way classification scheme provides a logical ordering and structure to the presentation and discussion of the job performance data. A summary description of the eight categories resulting from the use of this classification scheme is presented in Table 1.

Initially, the information will be presented in terms of the research conditions under which the data were collected. Subsequently, it will be discussed more extensively in terms of the data's relevance to major training issues:

- Transfer of training
- Simulation quality
- Effects of training on individual performance
- Differential effects of training methods
- Effects of training and experience on unit performance or operational readiness.

**TABLE 1. CLASSIFICATION SCHEME FOR CATEGORIZING
JOB PERFORMANCE MEASURES USED TO
ASSESS TRAINING EFFECTIVENESS**

1. SUBJECTIVE/PERSONNEL DIRECT/JOB PERFORMANCE DIRECT (S/PD/JPD)
Performance appraisal prepared by supervisors or technical experts who rate the effectiveness of job performance on the basis of direct observation or experience; e.g., supervisor's performance appraisal.
2. SUBJECTIVE/PERSONNEL DIRECT/JOB PERFORMANCE INDIRECT (S/PD/JPI)
Performance rating based on direct observations of performance on tests or job samples considered representative of the real maintenance tasks; e.g., Interview Troubleshooting Rating.
3. SUBJECTIVE/PERSONNEL INDIRECT/JOB PERFORMANCE DIRECT (S/PI/JPD)
Performance appraisal based upon group accomplishment; e.g., a unit rating or commendation.
4. SUBJECTIVE/PERSONNEL INDIRECT/JOB PERFORMANCE INDIRECT (S/PI/JPI)
Performance rating based upon the evaluator's perception of group accomplishment; e.g., testimonial of the value of training to organizational maintenance.
5. OBJECTIVE/PERSONNEL DIRECT/JOB PERFORMANCE DIRECT (O/PD/JPD)
Direct measurement of the quality or quantity of an individual's maintenance performance; e.g., an individual's average speed in completing a maintenance task.
6. OBJECTIVE/PERSONNEL DIRECT/JOB PERFORMANCE INDIRECT (O/PD/JPI)
Score or measurement based on individual performance in completing a hands-on test or selected work sample considered representative of real maintenance tasks; e.g., Walk-Through Performance Test.
7. OBJECTIVE/PERSONNEL INDIRECT/JOB PERFORMANCE DIRECT (O/PI/JPD)
Score or measurement based directly upon group performance on a task; e.g., maintenance management records of the Work Center hours used to complete a task.
8. OBJECTIVE/PERSONNEL INDIRECT/JOB PERFORMANCE INDIRECT (O/PI/JPI)
Score or measure that indicates the effect of maintenance performance on unit performance or accomplishment; e.g., flight sortie rate off a carrier.

A. OBSERVE AND DOCUMENT ACTUAL JOB PERFORMANCE

There are some practical, rather than conceptual, problems associated with directly observing job performance as a way of collecting data. First, one must identify the representative and critical tasks on which job performance data are required (a similar effort is needed to design training courses). Then, one must develop a way to measure quality of performance on these tasks on the job and devise ways of collecting the required data without contaminating the data (i.e., without influencing the way in which the job is performed in the presence of the observer). Then, one must locate some graduates of those courses and, using this method, measure performance on the job as these critical tasks occur during the period of observation. This is a valid approach, but it produces small amounts of data at a relatively high cost per observation.

B. OBSERVE PERFORMANCE ON JOB SAMPLES UNDER CONTROLLED CONDITIONS

Instead of waiting for critical tasks to occur naturally on a job, one can prepare equipment on which course graduates can be asked to perform critical tasks in a controlled environment (PD/JPI). Actual equipment, modified for test purposes, can be used to exhibit selected malfunctions, or simulated equipment can be designed to serve the same purpose. The use of simulators provides flexibility with respect to the number and variety of tasks that can be examined, as well as means for measuring human performance. Some costs are obviously incurred in developing and using the required test equipment, whether simulators or modified actual equipment.

A disadvantage of this approach is that the required data are not collected directly on the job but in a more or less artificial test environment. Another potential shortcoming is that data collected using simulators may be viewed as less credible than those collected on the job, in part because it is generally not known how well the tasks represent what is actually done on the job. The principal advantage of this approach is that data can be collected under well-controlled conditions on tasks selected in advance.

C. ASSESS THE EFFECTS OF TRAINING FROM INFORMATION IN EXISTING DATA BANKS

It is reasonable to believe that well-trained personnel perform better on their jobs than do less-well-trained personnel. If this is so, the effects of better training should be observable in such indicators as the amount of time needed to repair various types of equipment, the number of components removed as defective that are found to operate normally when tested later at a repair facility, performance in field exercises, and level of readiness reported for particular military units. The military services routinely collect many types of data needed to operate, manage, and support military groups and their equipment. The question, then, is whether the effects of training can be inferred from various types of management data being collected routinely by the military services. Since such management data are not being collected for purposes related directly to training, one might expect, at best, to observe only gross rather than specific effects that are present; this is a limitation. In areas where the impact of training can be detected and confirmed, the data banks might disclose trends over long periods of time and among a wide sample of organizations; use of such data does not intrude on an organization as does the on-site collection of data in an experiment; these are advantages.

II. DATA ON JOB PERFORMANCE

Among the ways of collecting job performance data relevant to training, we found no reports of direct observations of people actually doing their jobs. That this is a feasible procedure was demonstrated by Christensen (1949) over 35 years ago. He wished to determine systematically what aircraft navigators do. His data show the frequency of the various tasks performed by navigators, as observed every 30 minutes during flight. These data on job performance were used to design job aids and to improve cockpit layout but not to evaluate the effectiveness of training. The following sections consider job performance data collected under controlled conditions in the field, data from field survey studies, performance measurement and simulation quality, and data relevant to training derived from existing data banks.

A. JOB PERFORMANCE DATA COLLECTED UNDER CONTROLLED FIELD CONDITIONS

This category consists of data on job performance observed on selected tasks in a test environment near the job site rather than on actual tasks performed routinely. It is also necessary to know how those being observed were trained.

An unusual opportunity to measure the effectiveness of simulators as a way of training maintenance technicians was provided by the phased introduction of the F-16 Simulated Aircraft Maintenance Trainer (SAMT) by the U.S. Air Force in 1982. The SAMT is used by Field Training Detachments (FTDs) to provide additional maintenance training (after personnel have completed courses at technical training schools) on the specific aircraft equipments assigned to particular Air Force bases. Since the F-16 SAMTs were being introduced on a time-phased schedule, it appeared possible to compare the job performance of those trained on SAMTs with those trained conventionally on Actual Equipment Trainers (AETs). The method chosen was to measure the performance of these two groups on selected job samples.

1. Desired Maintenance Results

In one study, the Center for Competency Development (CCD, 1983) developed a rating form for supervisors to use in assessing the ability of maintenance technicians to diagnose discrepancies on the F-16 aircraft (S/PD/JPD). The form, called Desired Maintenance Results (DMR), provides explicit standards for rating job performance at levels ranging from unacceptable to perfect (1 to 6). Eight characteristics of work performance were evaluated:

1. Job Completion: quality, punctuality, and safety
2. Repeat/Recurrence: sortie abort, sortie delay, and loss of maintenance man-hours
3. Side Effects: new-problem and productivity loss
4. Resource Use: tools/equipment, spare parts, personnel, and expendables
5. Job Readiness: tools/equipment, expendables/spare parts, and personnel availability
6. Paperwork: reliability and efficiency
7. Housekeeping: litter, potential occupational and/or safety hazard, job cleanup, and appearance
8. Esprit de corps: "people friction," tardiness, and absenteeism.

A limited validation showed that the DMR ratings by supervisors produced the same results as ratings by subject matter experts (SMEs); i.e., there were no significant statistical differences between these two sources of ratings.

2. Troubleshooting Interview

A second endeavor by the Center for Competency Development (1983) involved the development and administration of a Troubleshooting Interview Rating which consisted of presenting one of two troubleshooting problems appropriate to the technician's AFSC (S/PD/JPI). The responses were graded by subject matter experts on a six-point scale. A rating of "1" is unacceptable performance, "3" is minimally acceptable, and "6" is perfect. The Troubleshooting Interview was presented to Weapons Specialists (462X0) and Flight Control Specialists (326X7) at two Air Force bases (AFB). One AFB used dedicated training devices for FTD training while the other used AETs. The results are shown in Table 2.

TABLE 2. TROUBLESHOOTING INTERVIEW RATINGS FOR
TD- AND AET-TRAINED TECHNICIANS^a

MAINTENANCE SPECIALTY	EXPERIENCE (MONTHS)	TRAINING METHOD	n	MEAN	t
Weapons	<6	HT	5	2.5	0
		AET	5	2.5	
	>6	HT	5	4.5	5.9*
		AET	5	2.8	
Flight Control	>6	SAMT	10	3.1	2.2*
		AET	7	3.6	

^aAdapted from CCD (1983), p. 33

*p < .05

The data for the Weapons Specialists indicate that the performance of the hardware trainer (HT) and AET groups with little work experience (less than 6 months on base) was essentially equivalent. Those trained with HTs who had more than 6 months of work experience after training had significantly higher Troubleshooting Ratings than those trained with AET: both groups improved with more time on the job, but the HT-trained group showed greater improvement.

Findings were different for the Flight Control Specialists, where the AET-trained group performed significantly better than the SAMT-trained group, although both were rated as only minimally acceptable.

The CCD report also presented data showing the effects of work experience on Troubleshooting Interview Ratings. [Because of differences in the levels of difficulty of the troubleshooting problems used by CCD, the data were converted to standard scores (mean = 3, sd = 1) and replotted in Fig. 1.] These cross-sectional data indicate that the first year after completing the FTD training is a period of rapid learning. A performance plateau seems to be reached after a year of experience. The dip at the end of Fig. 1 may represent the selective progression of the more able technicians to skill level 7, while the less able technicians remained at the skill level 5 classification.

Most of the Troubleshooting Interview Ratings were in the marginally acceptable range (2.5 to 3.0). However, based on the data presented, it is impossible to determine

whether the low scores were due to the quality of the training or to the level of difficulty of the questions.

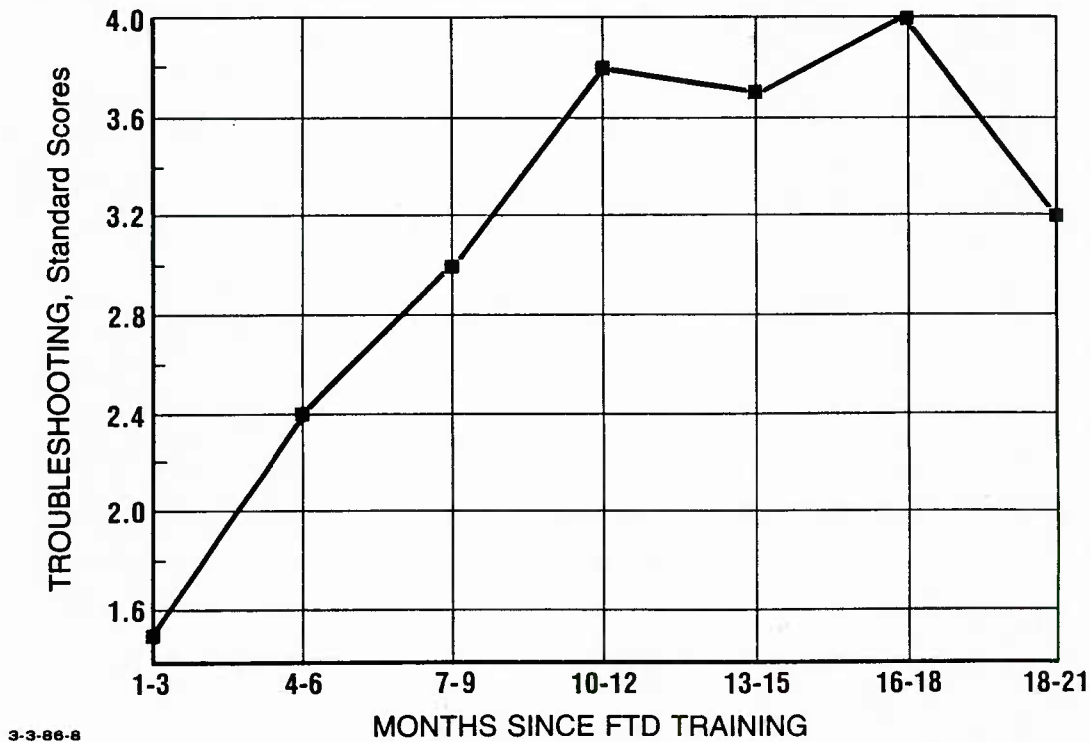


FIGURE 1. Troubleshooting performance as a function of experience^a

^aReplotted from CCD (1983)

3. Behaviorally Anchored Ratings

A different approach to the evaluation of maintenance training effectiveness was used in a field study of F-16 maintenance reported by Wienclaw and Orlansky (1983) and SAMTs were used for FTD training at Hill AFB in Utah and Hahn AFB in Germany. AET were used for FTD training at Nellis AFB in Nevada. In order to evaluate personnel performance both as students and technicians, a seven-factor behaviorally anchored rating scale was developed and used (S/PD/JPD). Both training and technician ratings were collected on all of the individuals involved in the study from their appropriate training and operational supervisors. Performance ratings were based on the following seven factors:

1. Safety
2. Thoroughness
3. Use of Technical Data
4. System Understanding
5. Understanding of Other Systems
6. Mechanical Skills
7. Attitude.

All correlations between the ratings of course graduates as students and as technicians are positive but low (see Table 3); they range from .11 on Understanding of Other Systems to .53 on Use of Technical Data, with a median value of .27. Thus, ratings during training account for 3 to 25 percent of the variance of the technicians' performance ratings. Only one of the correlations (Use of Technical Data) was statistically significant at the .05 level of confidence. The absence of statistically significant correlations, with one exception, may be due to the restricted range of the scores. Most of the scores on the rating scales, which had a range of 1 to 6, actually fell between 4 and 6. This would tend to reduce the magnitude of the correlations, and the small sample size ($n = 18$) requires higher correlations to achieve statistical significance. Since learning is not completed at the end of training, differential rates of learning and different absolute capabilities of the personnel could be expected to cause shifts in the relative positions of the individual rankings, with a resultant decrease in student and technician correlations.

TABLE 3. CORRELATIONS BETWEEN PERFORMANCE RATINGS (BARS) OF COURSE GRADUATES AS STUDENTS AND TECHNICIANS^a

<u>PERFORMANCE MEASURE</u>	<u>PEARSON r</u>
Safety	.22
Thoroughness	.17
Use of Technical Data	.53*
System Understanding	.38
Understanding of Other Systems	.11
Mechanical Skills	.16
Attitude	.33

^aWienclaw and Orlansky (1983), Table 3

* $P \leq .05$ ($N = 18$)

All of the BARS scores were higher for technicians than for students. Four of the seven comparisons were significant at the .05 level of confidence and all seven would be significant at the .10 level of confidence. The detailed statistical results are presented in Table 4. More detailed results are presented in Fig. 2, which shows a plot of the mean student and technician BARS scores for both the SAMT- and the AET-trained technicians for all seven scales. (BARS ratings were obtained from the technicians' supervisors and then from their training instructors, based on their memories of the same individuals as students. The average time interval between the completion of instruction and the rating as technicians was 3.5 months.) Several points are worth noting. The AET-trained personnel scored higher both as students and as technicians than did the SAMT-trained personnel. However, there is no way of determining whether this was a training difference or an inherent difference due either to chance or the selection and assignment process. The SAMT-trained personnel appear to have improved more than the AET-trained personnel. In general, the differences between the two groups were less for technicians than for students. Whether this is because the SAMT-trained technicians learned more from their operational experience or whether it is because the AET-trained personnel reached a ceiling in the ratings sooner cannot be determined from available data. It should be noted that the absolute differences in the scores were small and that the average ratings were high for both training groups (recall that average ratings noted in the CCD report were low).

TABLE 4. STATISTICAL EVALUATION OF IMPROVEMENTS IN BARS SCORES OVER TIME^a (F VALUES FOR REPEATED MEASURES ANOVA^b)

<u>PERFORMANCE MEASURE</u>	<u>F VALUE</u>	<u>p</u>
Safety	5.24	.036
Thoroughness	4.10	.060
Use of Technical Data	6.07	.026
System Understanding	4.16	.058
Understanding of Other Systems	3.22	.092
Mechanical Skills	6.26	.024
Attitude	13.92	.002

^aWienclaw and Orlansky (1983), Table 4

^bAnalysis of variance

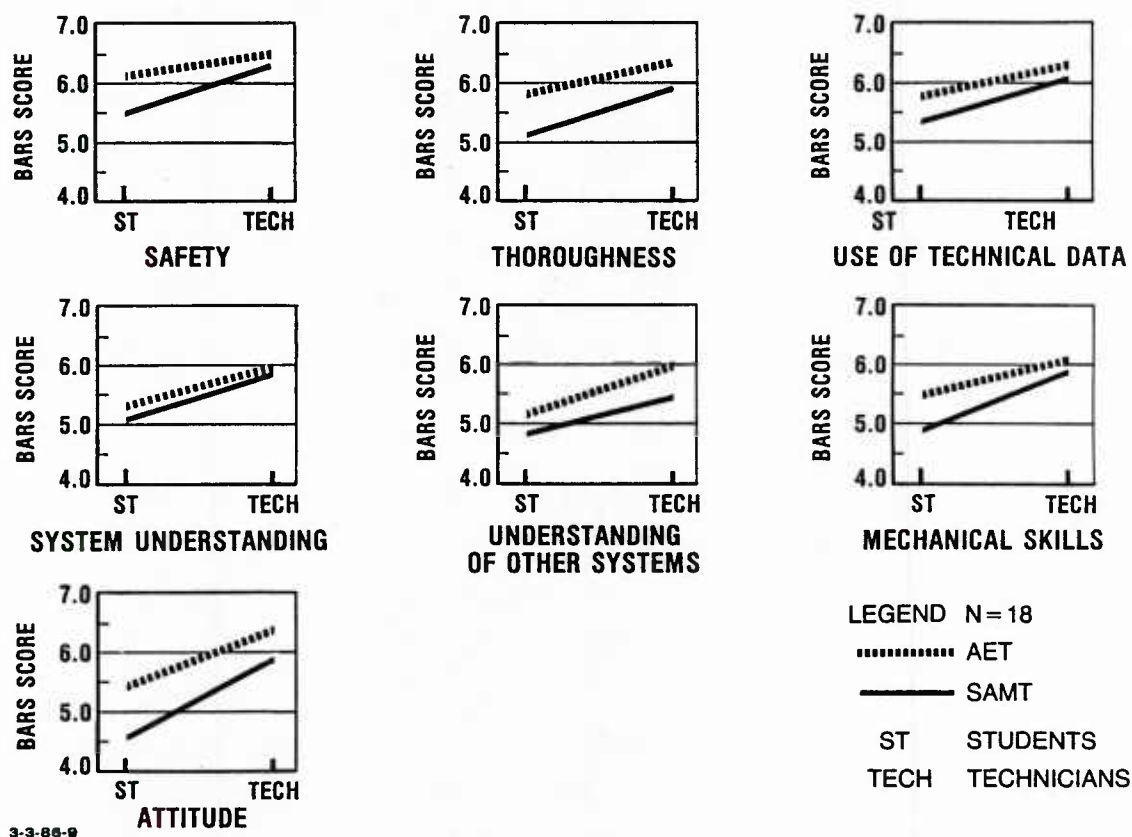


FIGURE 2. Performance ratings (BARS scores) for SAMT- and AET-trained students and technicians^a

^aWienclaw and Orlansky (1983)

4. Job Performance Measurement System

The Air Force has begun to develop the beginnings of an extensive Job Performance Measurement System (JPMS), Hedge, Ballentine, and Gould (1985). The system consists of eight hands-on performance tests (O/PD/JPI), seven interview tests (S/PD/JPI), and four rating forms (S/PD/JPD). Initial versions of the Walk Through Performance Test for TF-33 engine maintenance have been completed. Test data were collected from four Air Force bases. Test intercorrelations are presented in Table 5. The total WTPT scores correlated 0.96 with the hands-on performance tests and 0.60 with the interview tests. The hands-on tests correlated 0.44 with the interview tests. Training

grade, time in unit, remedial instruction, and mechanical aptitude score contributed significantly to a multiple R, predicting WTPT performance scores. The detailed multiple regression results are presented in Table 6. Ratings by supervisors correlated consistently (0.20 to 0.39) with total, hands-on, and interview scores of the WTPT. This represents a consistent but low level of common variance (4 percent to about 16 percent) between the supervisor ratings and the WTPT. Peer and self-completed versions of the rating forms showed no consistent relationship to the other WTPT measures. In a separate study, Hedge, Dickinson, and Bierstedt (1985) reported a WTPT test-retest reliability of .82 ($n = 12$, $p < .01$).

TABLE 5. INTERCORRELATIONS BETWEEN WALK-THROUGH PERFORMANCE TEST (WTPT) SCORES AND COMPONENT SUBTEST SCORES^a

	<u>WTPT</u>	<u>HANDS-ON</u>	<u>INTERVIEW</u>
WTPT	--	.96	.60*
Hands-On		--	.44*

^aAdapted from Hedge, Ballentine, and Gould (1985), Table 1

*Significant at the .05 level of confidence, $N = 84$.

TABLE 6. MULTIPLE REGRESSION OF PERFORMANCE AND TRAINING VARIABLES ON WTPT PERFORMANCE^a

<u>INDEPENDENT VARIABLES</u>	<u>MULTIPLE R</u>	<u>R SQUARE</u>	<u>R SQUARE CHANGE</u>	<u>SIMPLE R</u>	<u>B</u>	<u>BETA</u>
Training Grade	.28*	.08	.08	.28	.68	.38
Time in Unit	.36*	.13	.05	.26	.23	.22
Remedial Instruction	.41*	.17	.04	-.06	.69	.30
Mechanical Aptitude	.45*	.20	.03	.25	.15	.20
Time in Service	.47	.22	.02	.24	.01	.16
Time on Engine	.47	.22	.00	.18	-.11	-.10
Task Experience	.47	.22	.00	.13	.15	.01

^aHedge, Ballentine, and Gould

* $p < .05$

The JPMS represents an extensive effort to achieve a reasonable set of measures of maintenance performance that may serve as criterion measures for evaluating the effectiveness of selection and training procedures. The WTPT scores are sensitive to variations in a set of predictor variables: training grades, time in unit (experience), remedial instruction, and mechanical aptitude. Supervisor ratings have a low positive correlation with the WTPT scores; however, this is about the same order of magnitude of correlation typically found between selection and training scores and operational performance evaluations. At this point it is impossible to tell whether the tests really measure maintenance performance or simply the same test-taking abilities and general mechanical aptitudes measured by most selection and training tests. The ability to interpret the relevance of the WTPT scores to training would be greatly enhanced if they could be related to some other objective measures of the speed or quality of maintenance performance on the job.

B. JOB PERFORMANCE DATA COLLECTED UNDER FIELD SURVEY CONDITIONS

A number of interview, survey, and correlational techniques have been used in attempts to determine the relationship between training and operational performance. This category applies to job and job-related performance data collected routinely for management reasons. The investigators have attempted to relate these management data to the type of training or experience that the maintenance technicians involved in the study may have received. Maintenance management data banks contain voluminous amounts of data collected over relatively long periods of time.

1. Quality Assurance Personnel Test

One field survey effort conducted by Buchanan, Johnson, and McConnell (1982) endeavored to assess the impact of formal training at a Field Training Detachment (FTD) on the productivity of Air Force operational units. Technician performance was measured on the Quality Assurance Personnel Test (QA) (O/PD/JPI). Versions of the QA are administered routinely to maintenance technicians as a personnel quality control measure and to certify their ability to perform given levels and types of maintenance actions. Records of the QA scores are maintained routinely and are accessible for analysis. Data from three Air Force bases were collected: F-4 maintenance at George AFB; F-15

maintenance at Langley AFB and Luke AFB. Based on training records, the technicians were classified as trained, on-the-job (OJT) trained, or untrained. Trained personnel had completed the FTD training. OJT personnel had completed on-the-job training. Untrained had not completed either FTD or OJT training.

The results (see Table 7) were, at best, ambiguous. The F-15 QA data indicated that those who were FTD- or OJT-trained tended to do better than those who had not completed either training program. The F-4 data indicated that those who were FTD- or OJT-trained tended to do less well than those who had not completed either training program. Since it seems unlikely that training degrades an individual's ability to perform a technical task, there would appear to be a problem either with the classification procedures or the QA data or both. The definitions used to identify trained and untrained personnel, although intuitively appealing, seem to be inadequate. For example, many of the personnel at skill level 7 were listed as untrained--that is, without either FTD or OJT training. The achievement of a senior skill level without completing a training program would seem rather unlikely. Another problem with interpreting the data stems from the fact that the probability of passing the QA examination did not improve with skill level. The QA examinations appear to be tailored to the background and skill level. Consequently, these QA data do not seem to be very useful for assessing training effectiveness. However, the QA data could be very useful if they were accompanied by an index of the level of difficulty of the task which the technicians were performing.

TABLE 7. QUALITY ASSURANCE PERSONNEL TEST RESULTS^a

SOURCE	QA RESULT	TRAINING STATUS					
		TRAINED + OJT		UNTRAINED		TOTAL	
		n	%	n	%	n	%
F-15	Pass	1095	68	225	66	1320	68
	Fail	507	32	117	34	624	32
F-4	Pass	97	62	61	77	158	67
	Fail	60	38	18	23	78	33
Combined	Pass	1192	68	286	68	1478	68
	Fail	567	32	135	32	702	32

^a Adapted from Buchanan, Johnson, and McConnell, 1982, Exhibit III-2.

2. Task Completion Time

A second possible measure of maintenance performance was also evaluated in the same report by Buchanan et al. (1982). In this case, the time to complete a maintenance action was evaluated as a function of the level of training of personnel within a Work Center (O/PI/JPD). The data presented in Table 8 compares the ratios of the job completion times in the Work Centers with a high percentage of FTD-trained technicians to the completion times in the Work Centers with lower percentages of FTD-trained personnel. Comparison ratios were formed by dividing the job completion time of the Work Center with the higher percentage of FTD-trained personnel by the job completion time of the Work Center with the lower percentage of FTD-trained personnel. Ratios of less than one indicate that Work Centers with a higher percentage of trained personnel completed their maintenance tasks faster than Work Centers with a lower percentage of trained personnel. The data indicate that Work Centers with the higher percentages of FTD-trained personnel perform maintenance faster than Work Centers with a lesser percentage of trained personnel.

TABLE 8. TASK COMPLETION TIME RATIOS^a

COMPARISONS	SOURCES		
	F-4	F-15	Combined
Number of comparisons	79	136	215
Average percentage trained of more-trained Work Centers	55.5%	71.8%	68.4%
Average percentage trained of less-trained Work Centers	34.2%	40.7%	43.8%
Number of comparisons in which the Work Centers with higher percentages of FTD- trained had faster job completion times	46%	77%	123*
Average time ratio ^b	0.977	0.963	0.982

^a Adapted from Buchanan, Johnson, and McConnell, 1982, Exhibit III-5.

^b A ratio <1 indicates that the Work Centers with a higher proportion of FTD-trained personnel performed maintenance faster than Work Centers with a lesser proportion of FTD-trained personnel.

* Chi Square = 4.19, df = 1, p < .05.

The results of this study are of interest because they demonstrate that even relatively crude measures taken from maintenance management data banks are at least marginally sensitive to the effects of training on maintenance performance. More precise measures of maintenance performance inherent in these data banks (as will be shown later) can provide useful information on questions of training effectiveness.

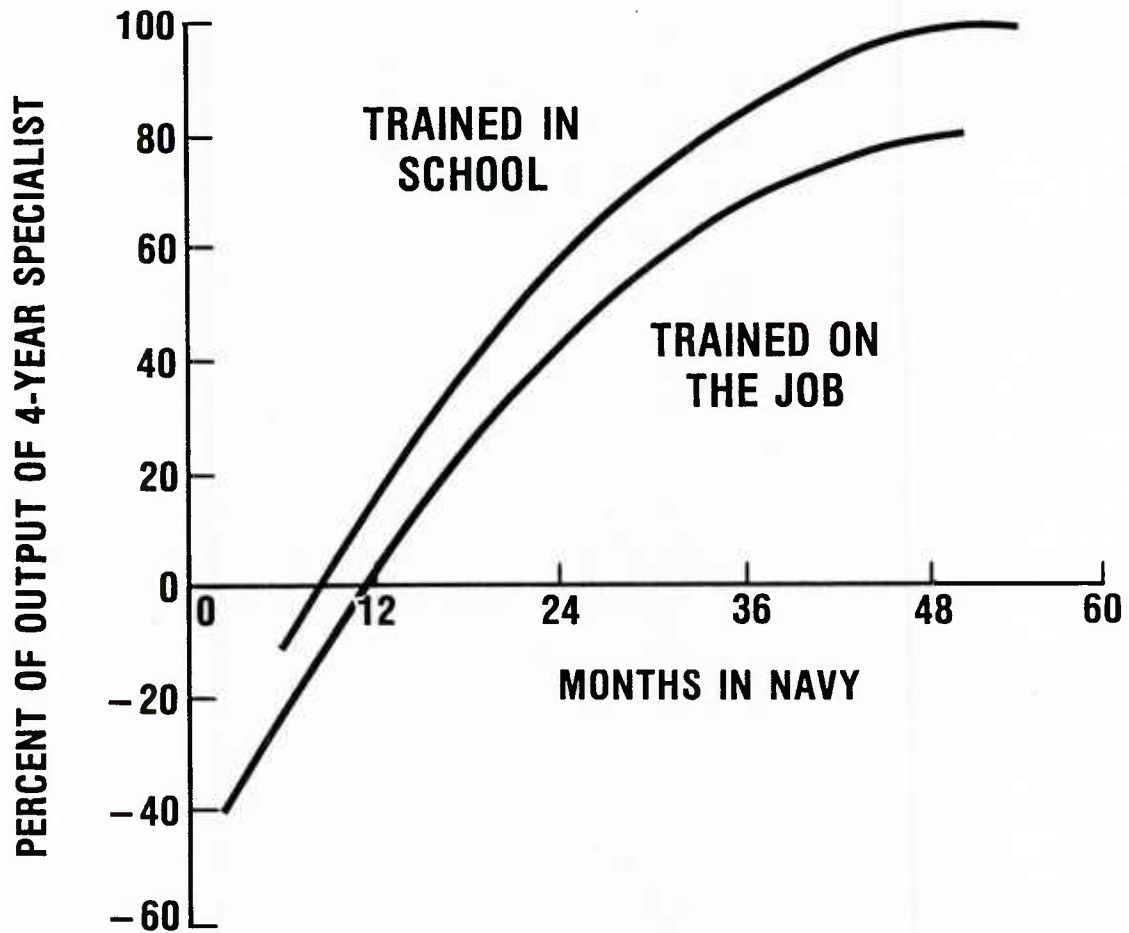
3. Net Productivity

Portions of the Enlisted Utilization Survey pertaining to Navy enlisted performance were analyzed by Quester and Marcus (1985). (The Enlisted Utilization Survey data were collected by Rand Corporation for the Defense Advanced Research Projects Agency.) Supervisors were requested to estimate the net effectiveness of personnel during four different time intervals within an initial 4-year enlistment period. Net productivity was the estimate of an individual's productivity minus any supervisory time required to achieve that level of performance, compared to the output of a specialist trained for four years ("100 percent"). Two thousand supervisor estimates involving 15 Navy specialties were collected between November 1974 and January 1975. Net productivity estimates were made for four time intervals: 1 month, 1 year, 2 years, and 4 years.

For all occupations measured, the average level of productivity increases over time. Figure 3 presents the supervisors' estimates of the productivity growth for electricians' mates in the first enlistment; this figure is representative of the principal results of the analysis. In occupations that offer alternative training paths, the productivity of the A School graduates exceeds that of those learning exclusively on the job. The A School graduates were significantly more productive at each of the four rating points. The typical OJT trainee never reaches the level of the "average four year specialist." Average productivity after 4 years at the duty station is approximately 100 percent for A School graduates.

The estimate of net productivity appears to be a very useful rating scale. The net productivity measures were sensitive to differences between training methods: A School or OJT. On the negative side, there is a hazard that the data may reflect supervisory opinions about the benefits of A School training rather than real differences in performance. There is also a possibility that there may be differential selection involved in the assignment of personnel to OJT or A School training, in which case the differences may be largely the result of personnel differences rather than differences in training methods. More

positively, the net productivity measure has the advantage of being a universal measure which is quick, simple, and easily interpretable: this is a property that some of the more specific performance test measures lack.



3-3-86-10

FIGURE 3. Productivity growth for electricians' mates in the first-term enlistment^a

^aQuester and Marcus (1985)

C. USE OF SIMULATORS TO COLLECT JOB PERFORMANCE DATA

There are many advantages to the use of simulated equipment as a way of collecting job performance data; e.g., convenience, easy access to a wide variety of operating and equipment conditions, and ease of measuring the performance of personnel. Nevertheless, it is reasonable to ask whether the quality of data collected on job performance is influenced by the quality of the simulator.

The magnitude of the transfer of knowledge and skills learned in a simulator to performance on the job should be related to the degree of physical and behavioral correspondence between the tasks performed in the simulator and the tasks performed on the job. Experience in training should establish an intellectual and performance readiness base such that personnel can gain rapidly in competence from their operational experience.

1. Physical and Behavioral Correspondence

Jorna and Moraal (1985) report a series of comparisons between performance in a simulator and performance with actual equipment for both students and experienced personnel. This type of approach might serve as a beginning model for assessing the behavioral fidelity of simulators (O/PD/JPD). Although this study analyzes performance in a tank-driving simulator, the concepts involved should be applicable to objective evaluations of the physical and behavioral fidelity of maintenance-training simulators.

The training time required to reach criterion levels of performance for four tasks was compared for students trained in tanks and in simulators. The results are presented in Fig. 4. The required training time is a function both of the task and the method of training. Gear-changing was learned faster in the simulator, but the steering task took longer to learn. Little additional training time was required to perform at criterion levels in the tank after training in the simulator. The performance of experienced individuals was also measured in both the tank and the simulator. Mean performance values of the experienced drivers are presented in Table 9. If there is a high level of behavioral fidelity, the experienced drivers would be expected to demonstrate similar performance in both the tank and the simulator and there would be no learning on successive trials in the simulator. The gear-changing task would meet the criterion of high behavioral fidelity. The steering task resulted in lower performance in the simulator than in the tank and interacted with the combination tasks to produce lower performance and a learning effect on successive trials.

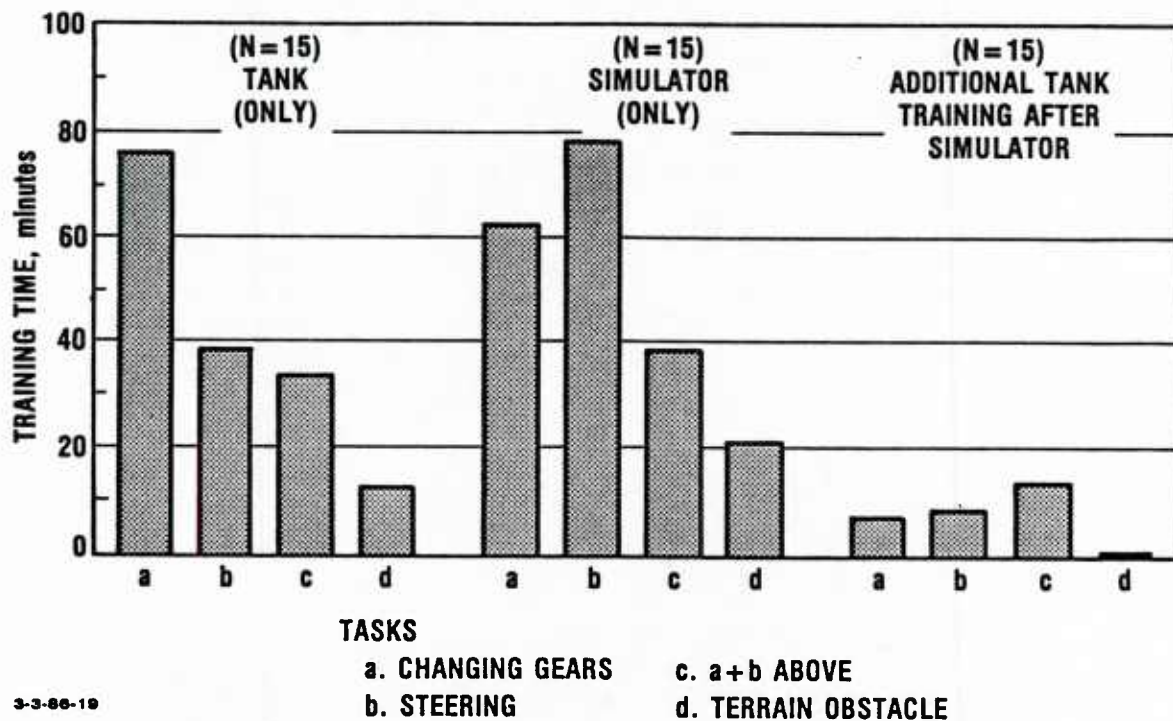


FIGURE 4. Mean training times needed by groups trained in a tank or simulator to reach criterion performance^a

^aJorna and Moraal (1985)

TABLE 9. MEAN VALUES OF EXPERIENCED TANK DRIVERS ON FOUR TASKS^{a,b}

TASKS	TANK	SIMULATOR (1st)	SIMULATOR (2nd)
(a) Changing gears:			
Accelerating (s)	27.3	28.3	28.5
Decelerating (s)	21.2	19.5	19.4
(b) Steering			
Number of errors ^c	1.4	3.9	3.8
(c) Time to complete trajectory (s)	37.7	34.6	32.9
(d) Terrain obstacle	0.4	5.7	3.7

^aJorna and Moraal (1985)

^bn = 10

^cNumber of pylons hit

These results are significant for several reasons. First, they provide a model for evaluating the behavioral fidelity of the simulator when compared with the actual equipment. Second, they demonstrate the importance of being able to evaluate fidelity at the task level. Third, they demonstrate that fidelity can be assessed with objective data.

2. Instructor Ratings of Simulator Fidelity

Fitzpatrick and Hritz (1984) used F-16 SAMTs to study the effects of simulator fidelity on student performance. They compared student confidence and task performance error rates with instructor ratings of simulator fidelity. Six instructors rated the comparative fidelity of four SAMTs:

1. TFE-2 Flight Control/Instrumentation
2. TFE-4 Electronics
3. TFE-11 Engine Diagnostics
4. TFE-12 Engine Operating Procedure.

(A detailed list of the F-16 simulators and training devices is presented in Table 10.) The instructors first rated the overall fidelity of the simulators as "High," "Middle," or "Low." They then rated the comparative fidelity of operational checks and fault isolation checks within each trainer. Instructor ratings, student confidence ratings and performance errors are summarized in Tables 11 and 12. Students' confidence levels in performing end-of-course tasks are presented in Fig. 5. The proportion of end-of-course errors in performing maintenance tasks is presented in Fig. 6. (Figures 5 and 6 were replotted from data provided in Fitzpatrick and Hritz, 1985, Figs. 1 and 5.) The instructors rated the Engine Operating Procedure (Run) Trainer as having the highest fidelity and the Flight Control/Instrumentation Trainer as having the lowest fidelity of the four simulators.

In general, student confidence ratings (S/PD/JPI) and end-of-course performance error scores (O/PD/JPI) were consistent with the instructors' ratings. The only exception was the comparative fidelity ratings of operational checks and fault-isolation checks on the Engine Operating Procedures Simulator. Instructors rated the fault isolation checks as being better than the operational checks, but students performed better on the operational checks. However, the difference was small and the overall performance on the Engine Operating Procedures Trainer was notably better than on the other simulators.

TABLE 10. F-16 SIMULATORS AND TRAINING DEVICES

NUMBER-NAME-TYPE (Manufacturer)	WUC ^a -EQUIPMENT	COURSES-AFSC ^b
TFE-2--Flight Control. Instrumentation SAMT (Honeywell)	14A00--Primary Flight Control Electronics 14B00--Primary Flight Control Actuators 51A00--Primary Flight	Integrated Avionics-- Instrument and Flight Control System Specialist (F-16)-- AFSC 326X7
TFE-3--Navigation SAMT (Honeywell)	71A00--TACAN Navigation Set 71B00--Instrument Landing Set	Integrated Avionics--Navigation and Penetration Aids Systems Specialist (F-16)--AFSC 326X8
TFE-4--Electronics SAMT (Honeywell)	42000--Electrical Power Supply	Aircraft Electrical Systems Tech- nician (F-16)--AFSC-423X0
TFE-6--Seat and Canopy-- Hardware Trainer (General Dynamics)	12000--Crew Station System	Aircrew Egress Systems Tech- nician (F-16)--AFSC 423X0
TFE-10--Engine Start SAMT (Honeywell)		
TFE-11--Engine Diagnostics SAMT (Honeywell)	23000--Turbofan Power Plant 24000--Auxiliary Power Plant	Jet Engine Technician (F-16)-- AFSC 426X4
TFE-12--Engine Operating Procedure SAMT (Honeywell)		
TFE-13--F-100 Engine Hardware Trainer (General Dynamics)		
TFE-14--Gun-- Hardware Trainer (General Dynamics)	75A00--Gun System	Weapons System Maintenance Technician (F-16)-- AFSC 462X0
TFE-15--Fuel-- Hardware Trainer (General Dynamics)	46000--Fuel System	Aircraft Fuel Systems Technician (F-16)-- AFSC 423X3
TFE-22--Environmental Control--SAMT (ECC)	41000--Environmental Control System	Aircraft Environmental System Technician (F-16)-- AFSC 423X1

^a Work Unit Code.

^b Air Force Specialty Code.

TABLE 11. SUMMARY OF INSTRUCTOR RATINGS OF SIMULATOR FIDELITY AND STUDENT CONFIDENCE RATINGS AND PERFORMANCE ERRORS^a

SIMULATOR	INSTRUCTOR RATING (n = 6)	STUDENT		
		n	CONFIDENCE ^b	ERRORS
Flight Control	Low	11	4.0	12%
Electronics	Middle	11	5.6	4%
Engine Diagnostics	Middle	13	4.4	6%
Engine Operation/Run	High	20	5.8	3%

^a Adapted from Fitzpatrick and Hritz, 1984, Figures 1 and 2.

^b Confidence scale ranged from a low of 1 to a high of 6.

TABLE 12. SUMMARY OF INSTRUCTOR RATINGS OF THE COMPARATIVE FIDELITY OF OPERATIONAL CHECKS AND FAULT ISOLATION CHECKS AND STUDENT CONFIDENCE RATINGS AND PERFORMANCE ERRORS^a

SIMULATOR	INSTRUCTOR RATING (n = 6)		STUDENT				
			n	CONFIDENCE ^b		ERRORS	
	Ops	Fault		Ops	Fault	Ops	Fault
Flight Control	1st	2nd	11	3.9	4.0	6%	17%
Electronics	1st	2nd	11	5.9	5.2	1%	8%
Engine Diagnostics	2nd	1st	13	4.3	4.5	7%	5%
Engine Operation/Run	2nd	1st	20	5.6	5.9	2%	4%*

^a Adapted from Fitzpatrick and Hritz (1984), Figures 1 and 2.

^b Confidence scale ranged from a low of 1 to a high of 6.

* Only reversal from expected performance.

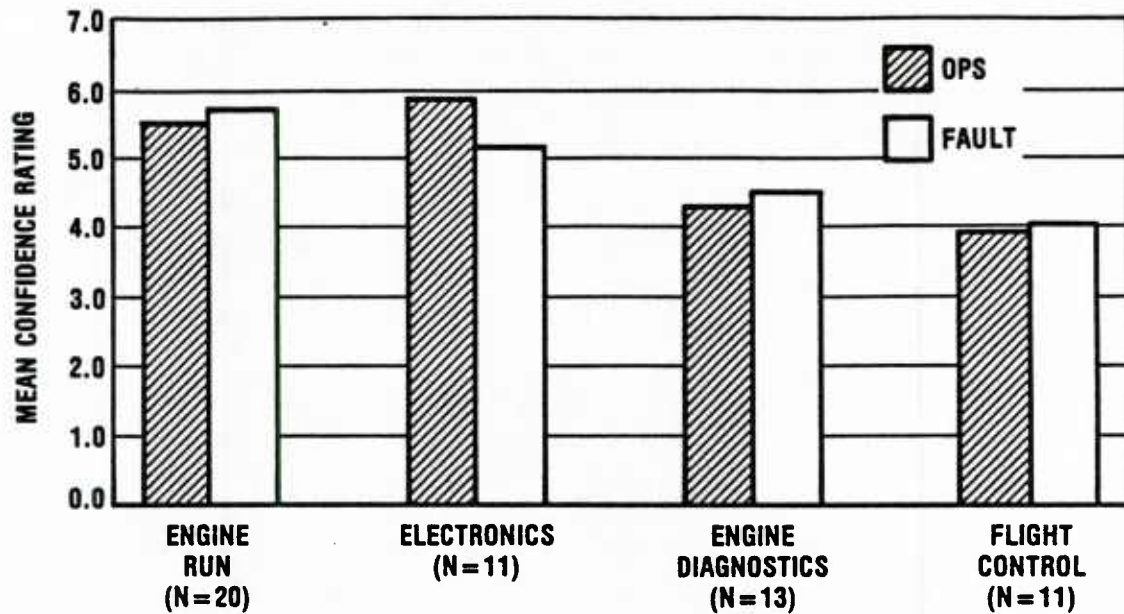


FIGURE 5. Student confidence in their ability to perform tasks on the simulator at end of course^a

^aFitzpatrick and Hritz (1984)

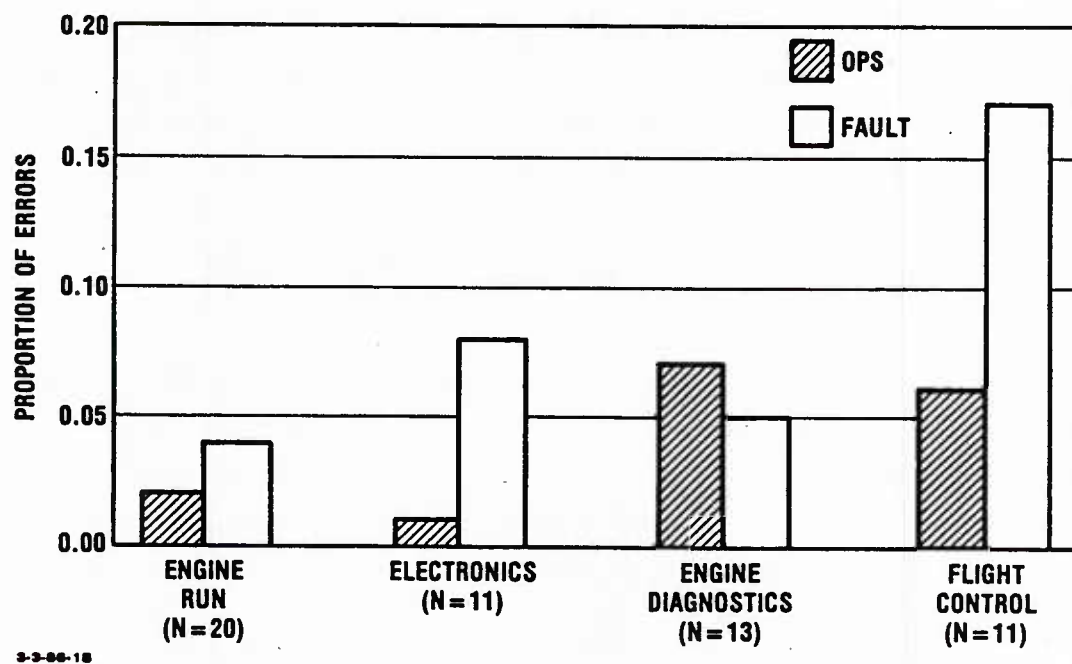


FIGURE 6. Proportion of errors on tasks performed on simulator at end of course^a

^aFitzpatrick and Hritz (1984)

End-of-course measurements of students' confidence ratings and actual performance of selected tasks on the simulator were related closely to instructors' ratings of simulator fidelity. Consequently, it can be concluded that both student confidence and performance are strongly influenced by the fidelity of the simulation. The study reinforces the concept that all simulators are not equally effective and that all tasks within a given simulator are not represented equally well.

D. DATA RELEVANT TO TRAINING DERIVED FROM EXISTING DATA BANKS

The ability to observe the effects of training and detect differences due to alternative methods of training in existing maintenance management data banks has a number of ideal properties. Each military service operates a large maintenance data bank. The usefulness of these data banks to yield data relevant to training has been examined (see String and Orlansky, 1981). One advantage to using information from these data banks is that it provides an objective measure of real maintenance performance. Like coins minted from precious metals, the data have inherent value and meaning while the value and meaning of the other types of objective measures usually depend on the closeness of their relationship to some acceptable criterion. Additionally, since data bank information is collected routinely and unobtrusively, it represents actual performance as opposed to data collected under test conditions that may be subject to the "Hawthorne effect." The data are meaningful to both managers and researchers, they can be used to track long-term trends, and their collection is a normal part of organizational management that does not disrupt normal activities.

Differences in the effectiveness of alternative methods of training should be manifest in a data bank in several ways. If one training method is more effective than another, the effects of the better method should be evidenced by better quality work or speed of work. Generally, better trained and more experienced personnel are faster than their less trained or less experienced counterparts. Finally, the effects of training and experience should ultimately be related to unit performance or combat readiness.

1. Quality of Work

Several types of routine measures have the potential value for assessing the effects of training on performance. Because equipment maintenance involves the detection,

identification, removal, and replacement of defective components, the accuracy and speed with which defective components are removed and replaced would provide measures of the quality and quantity of maintenance. A major component in evaluating combat readiness is the presence of major systems malfunctions. Training should manifest its effects through a decrease in the frequency of major system malfunctions. In aviation, system malfunctions result in a decrease in the number of flight sorties; consequently, an organization with better trained personnel should have more sorties than one which has maintenance personnel with less training and experience.

An earlier review of the performance of maintenance technicians by Orlansky and String (1981) indicated that, across a group of seven studies, non-faulty parts were removed in 4 to 43 percent of all corrective maintenance actions. Components removed by an operational organization (e.g., a flight squadron) are usually sent to another organization for repair, but before any repairs are made, the components are usually retested. Non-faulty parts are those that were removed but found not to be defective when received for repair.

A study by McConnell and Johnson (1984) on productivity in Air Force F-16 units sought to use data on Retest-OK (RETOK) rates as a measure of maintenance quality. They collected data from five Air Force F-16 wings, of which four used SAMTs and one used AETs for FTD training, but the effort proved unproductive. A set of management and operational practices resulted in an absence of usable data because (1) "Retest-OK" data were kept only when the rates exceeded 8 percent for the entire system, (2) many of the components had no turn-in tags, (3) many components are used on more than one aircraft, and (4) in most instances the wing or base, but not the Work Center, could be identified.

This McConnell and Johnson study highlights some of the problems in trying to use data from management data banks to evaluate training effectiveness. Data may not be acquired or kept in a form useful for training evaluation purposes. To be useful for evaluating the effect of training, it is necessary to be able to clearly relate the maintenance data to the specific system, Work Center, and if possible, the performing technician.

The number of major equipment malfunctions, or conversely, the absence of serious mission-degrading equipment failures are data-base measures which should vary with the quality of training and experience of the maintenance force. Several studies have found a positive relationship between equipment status and the training and experience of

maintenance personnel. Using Navy Casualty Reports (CASREPs) as their data source, Horowitz and Sherman (1977) reported that ships experience fewer major equipment problems when more experienced personnel are aboard. In another study, Horowitz and Angier (1985) reported several relationships between operational measures and the training and experience of the maintenance personnel. First, the fraction of surface combatant ships with no serious mission-degrading equipment failures (O/PI/JPI) between 1977 and 1983 varied as a function of the ratio of the number of junior (E-1 to E-4) personnel to the number of authorized billets and the ratio of senior (E-5 to E-9) personnel to the number of billets. Using regression analysis, they found that changes in the fill rate for senior enlisted maintainers were statistically significant and much more important than changes in the fill rate for junior personnel. Second, after reviewing the CASREPs for 91 ships over a 3-year period, they concluded that the experience level of the maintainers is the most consistent predictor of readiness (O/PI/JPI). Third, using the number of A-7 flights off a carrier in a quarter as a measure (O/PI/JPI), they concluded that adding one junior person (E-1 to E-4) to a ship seemed to depress performance, presumably because more of the time of the senior personnel was diverted to direct supervision. In general, the regression analysis data presented in Table 13 indicate that the presence of more senior personnel enhances operational performance.

Clearly, formal training and experience-induced training have an observable and meaningful impact on some operational measures. Refinement and more general use of these measures could provide a valuable source of data for assessing training effectiveness.

TABLE 13. MARGINAL PRODUCT OF PAYGRADE GROUPS IN GENERATING A-7 SORTIES PER QUARTER^a

PAYGRADE		
<u>E-1 to E-4</u>	<u>E-5 to E-6</u>	<u>E-7 to E-9</u>
-0.5	6.2	29.1

^aHorowitz and Angier (1985)

2. Speed of Work

As personnel gain familiarity and experience with a repair task, their speed in accomplishing the task increases; consequently, the time needed to accomplish a repair task should be related to the amount of training and experience of the maintenance personnel. Two studies have successfully used speed of work as a measure of training effectiveness which differentiates between two methods of training: SAMTs or AETs.

Using data from the Air Force Consolidated Data System for three F-16 wings for calendar year 1982, Johnson, McConnell, and Murdock (1983) found that speed in accomplishing maintenance tasks was related to the completion of FTD training. The measure of productivity used was the elapsed time per worker (O/PI/JPD). The data collected were limited to Work Unit Codes (WUCs) which had a simulator training option. Work Centers with a high percentage of FTD-trained personnel (60 percent or over) were compared with Work Centers with less than 60 percent of FTD-trained personnel. Performance data on two WUCs are presented in Figs. 7 and 8: WUC 23Z00, Turbofan Power Plant (F-100 engine) and WUC 14A00, Instrument and Flight Control Systems.

For both of the WUCs examined, FTD training had a greater effect on reducing the time needed to perform maintenance than did experience (i.e., the frequency with which the task was performed). The effect of training was statistically significant for WUC 23Z00, Turbofan Power Plant, and was present but not statistically significant for WUC 14A00, Instrument and Flight Control Systems. For both WUCs, there was an interaction between training and experience such that Work Centers with a high percentage of FTD-trained personnel exhibit a clear increase in productivity with increased workload. Conversely, Work Centers with lower percentages of trained personnel seemed to show decreased productivity with increased experience/workload.

Productivity, as measured by the elapsed time per completed work action, was sensitive to the effects of training in terms of the relative percentages of FTD-trained technicians within the Work Centers. The interaction between training and experience/workload suggests that those who have had the FTD training benefit or learn from the increased experience gained at higher workloads. This interaction might be labeled as an "experience readiness" effect.

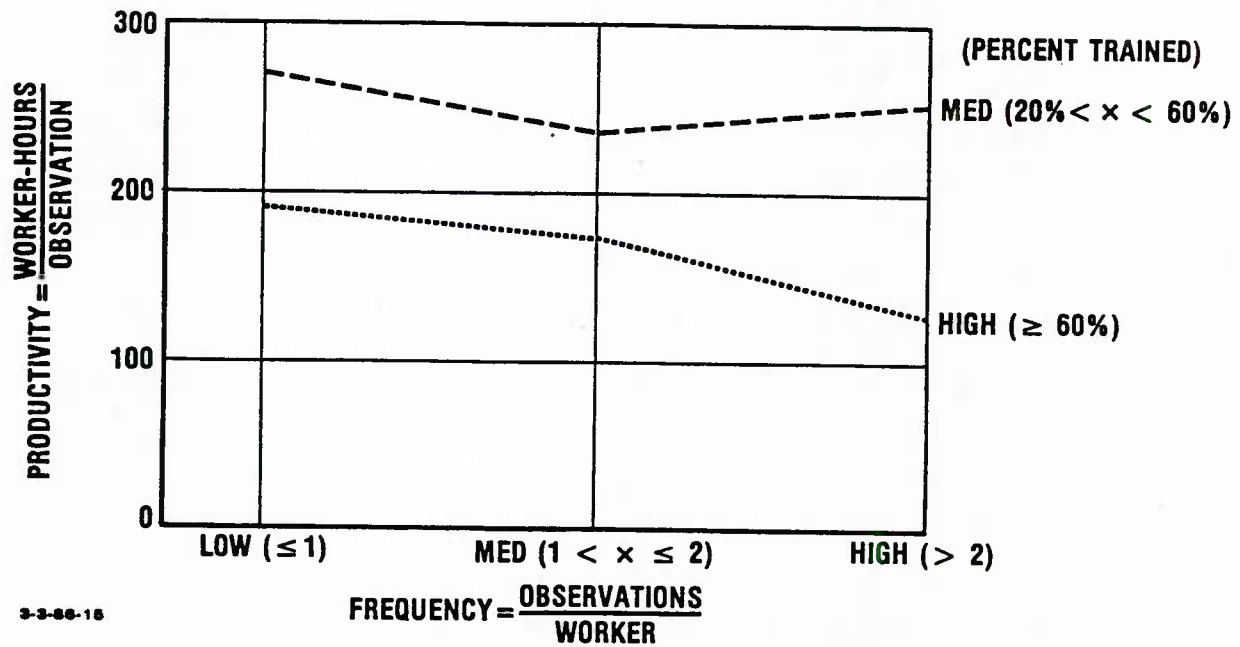


FIGURE 7. Effects of training on turbobfan power plant (WUC 23Z00) maintenance^a

^aJohnson, McConnell and Murdock (1983)

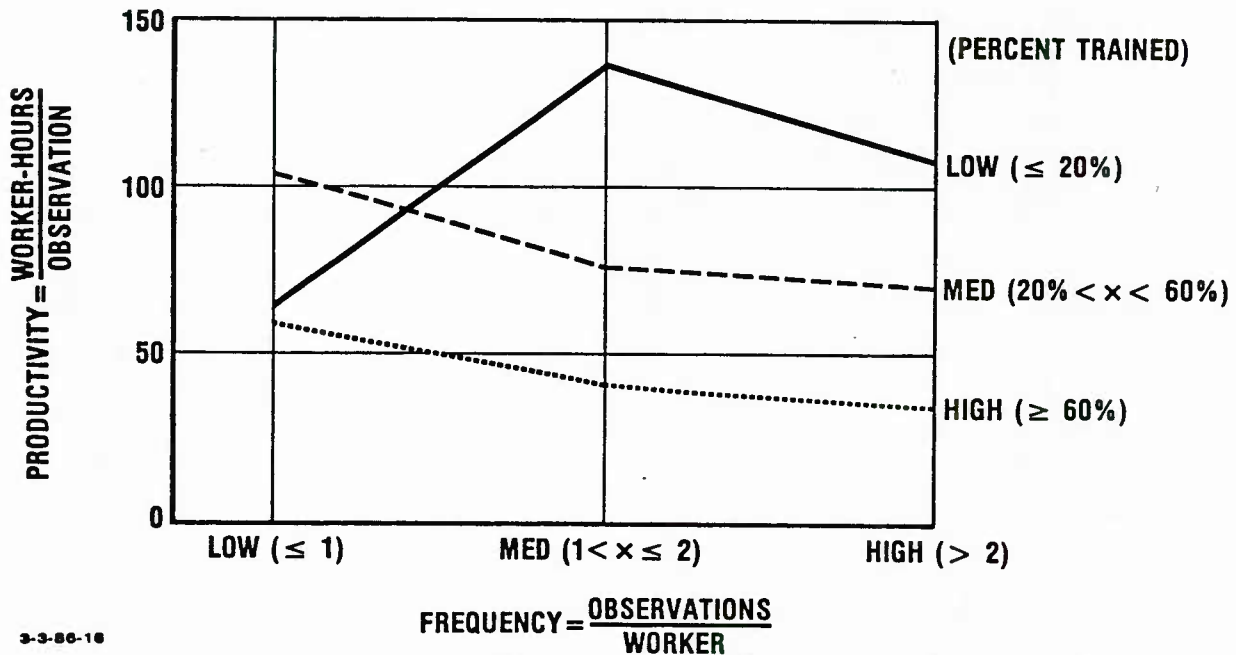


FIGURE 8. Effects of training on instrument and flight control systems (WUC 14A00) maintenance^a

^aJohnson, McConnell and Murdock (1983)

In a subsequent study, McConnell and Johnson (1984) were able to confirm and extend their previously reported findings. They collected data on five Air Force F-16 wings using information obtained from the Maintenance Data Collection (MDC) and the Maintenance Management Information and Control System (MMICS) for the first 6 months of 1983. The measure of productivity was work hours to complete a specific work action (O/PI/JPD). The unit of comparison was the Work Center, consolidated within and between wings for three WUCs:

- (1) Jet Engine (23000)
- (2) Aircraft Electrical Systems (42000)
- (3) Flight Control Systems (14000).

The results provided information on the effects of FTD training on productivity and the effects of using either SAMTs or AETs for the FTD training on productivity. The training effects vary with WUC, i.e., from system to system; therefore, we will review the specific effects on a system-by-system basis.

Maintenance productivity for WUC 23000, Turbofan Power Plant (AFSC 426X4), was related positively to the increasing percentage of FTD-trained personnel within the Work Center (see Fig. 9). Productivity was not related significantly to the frequency of performing a given task. There is an interaction between training and experience such that Work Centers with a high proportion of FTD-trained personnel do markedly better under high frequency conditions, while the Work Centers with lower proportions of FTD-trained personnel experienced maximum productivity under the low-to-medium frequency/workload conditions. The FTD-trained personnel seem to benefit more from work experience than those without FTD training. This appears to be another manifestation of an "experience readiness" factor.

Maintenance productivity for WUC 42000, Electrical Power Supply (AFSC 423X0), was related positively to the percentage of FTD-trained personnel within the Work Center (see Fig. 10). The frequency of performing a given task was not significantly related to productivity.

Maintenance productivity for WUC 14000, Flight Control Systems (AFSC 326X7) was not related to either the percentage of FTD-trained personnel within the Work Center or to the frequency of performing a specific task (see Fig. 11).

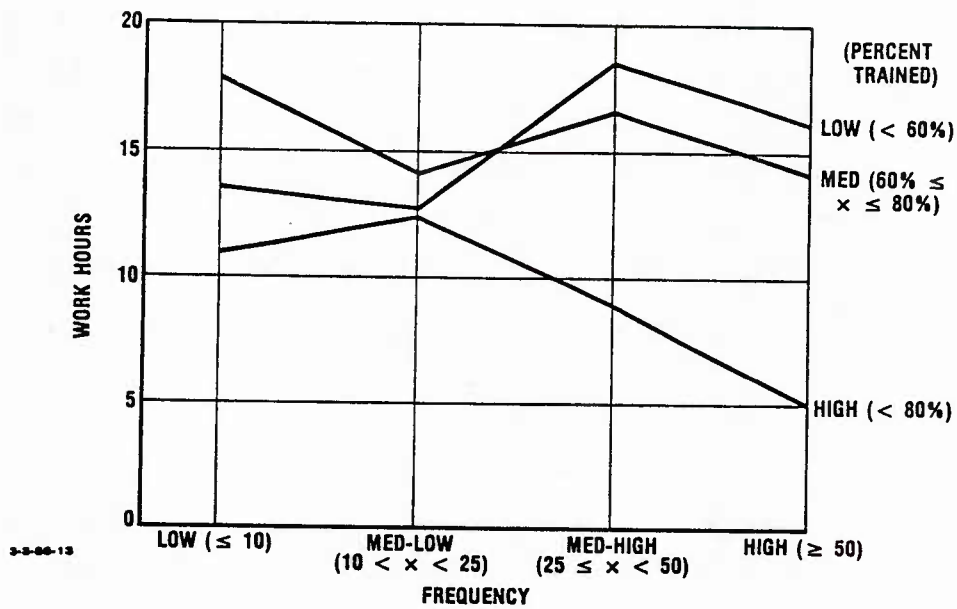


FIGURE 9. Aggregate training effects on turbofan power plant (WUC 23000) maintenance productivity^a

^aMcConnell and Johnson (1984)

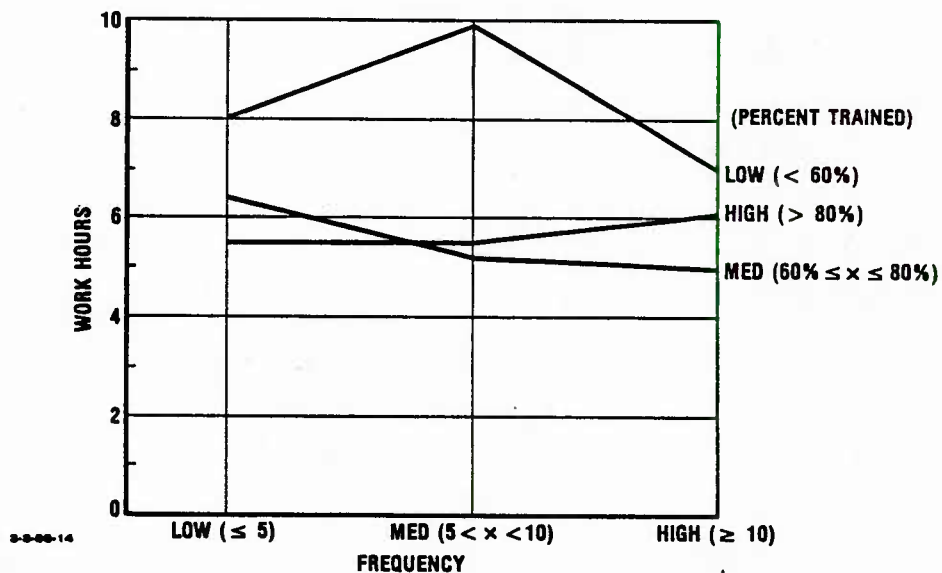


FIGURE 10. Aggregate training effects on electrical power supply (WUC 42000) maintenance productivity^a

^aMcConnell and Johnson (1984)

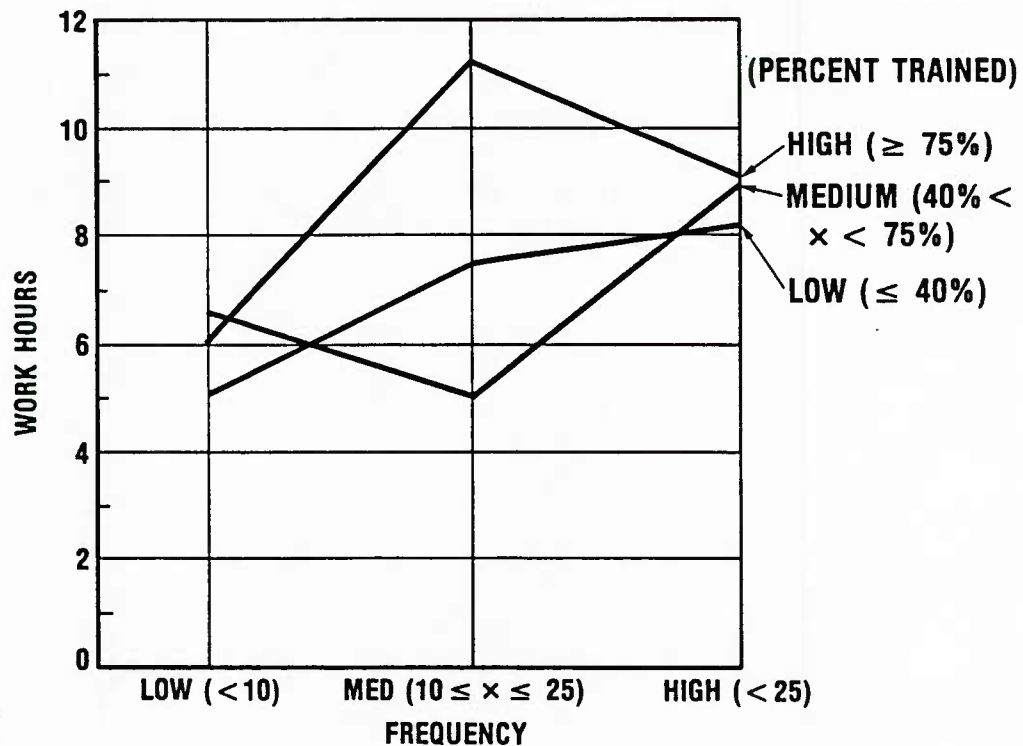


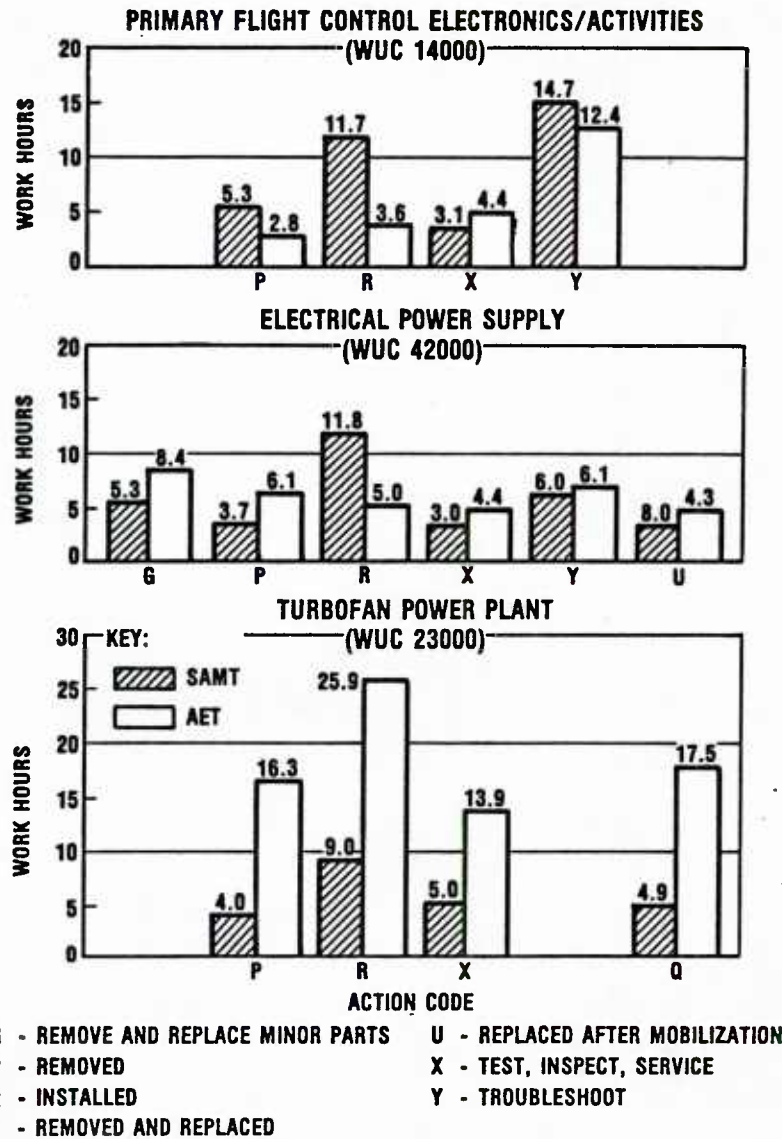
FIGURE 11. Aggregate training effects on flight control systems (WUC 14000) maintenance productivity^a

^aMcConnell and Johnson (1984)

The second portion of McConnell and Johnson's results provide us with a detailed comparison of the effects of SAMT and AET training on maintenance productivity. Four of the five air bases involved in the study used SAMTs for FTD training; one used AETs. A series of detailed comparisons was made between the productivity measures collected from the Work Centers of a wing which used SAMTs for FTD training (Luke AFB) and another which used AETs (Nellis AFB). The data contain a confounding factor of importance which needs to be considered in evaluating the results. The percentage of personnel completing FTD training was higher for those who used SAMTs than for those who used AETs, 89 percent compared to 66 percent, respectively.

The composite totals show that the personnel with SAMT training performed the maintenance tasks faster than their counterparts with AET training. As with the previous

results, in evaluating the effects of the presence or absence of FTD training on productivity, the specific results vary from system to system. A detailed presentation of the effects of FTD training methods on the time needed to complete specific maintenance tasks is presented in Fig. 12. The productivity data for WUC 23000, Turbofan Power Plant, show



^aReplotted from McConnell and Johnson (1984).
 3-3-88-11

FIGURE 12. Effects of FTD training methods on time needed to complete specific maintenance tasks^a

^aReplotted from McConnell and Johnson (1984)

that the SAMT-trained personnel are considerably faster than the AET-trained personnel on all four tasks on which a comparison can be made. (Since FTD training has been shown to improve performance, some unknown proportion of the difference between these two groups is probably due to differences in the percentages of FTD training rather than being due to the method of training.) Productivity data for WUC 42000, Electrical Power Supply, show SAMT-trained personnel as being faster than AET-trained personnel on five out of six comparison tasks. Even though the differences in the percentages of FTD-trained personnel would tend to favor the SAMT-trained group, it would still seem likely that the SAMT-trained technicians are at least equal to their AET counterparts. The productivity data for WUC 14000, Flight Control, show the AET-trained personnel as being faster on three out of four comparison tasks. In this instance, the differences in the percentages of FTD-trained personnel would tend to support the conclusion that the observed differences are real.

Work hours used to complete a task is a meaningful and useful productivity measure that is sensitive to differences in training backgrounds and methods. The percentage of personnel in an AFSC that are FTD-trained has a significant effect on Work Center productivity. The relative effectiveness of SAMT or AET training seems to vary from system to system. Most importantly, this study demonstrates the potential worth of objective measures of job performance (O/PI/JPD) as a means of measuring training effectiveness.

3. Combat Readiness

Several investigators have explored the possibility of using reports of unit combat readiness as a measure of training effectiveness. Pellicci (1985) describes a training readiness model which from any combat readiness level specifies the amount of additional training and resources needed to achieve full combat-ready status; however, the model is in the early stages of development and there is as yet no data available to validate its predictions on the relationship between the use of resources and the quality of combat readiness. Cavalluzzo (1985) has used the Training Readiness-Index Score (CRTRNG) (O/PI/JPI) contained in the Navy's Unit Status and Identity Reports (UNITREP) to evaluate factors related to a ship achieving full combat readiness. She found that the tempo of operations is strongly associated with the level of training readiness upon deployment.

An increase of 1-day-per-quarter in training was associated with a 2.26 percent rise in the number of ships that are combat ready upon deployment.

There appear to be a number of data base measures which show the effects of training, training methods, and experience-induced training on the quality and speed of maintenance. These measures are important because they are the technical criterion data needed to evaluate training methods and to validate other selection and training assessment techniques. The most effective ones were the time-needed-to-complete-specific-maintenance-tasks data reported by McConnell and Johnson (1984), which provided the basis for a detailed comparison of the training effectiveness of SAMTs and AETs for three Work Unit Codes. Data base measures are also important because they represent a necessary beginning in the process of establishing a meaningful quantitative linkage between maintenance training and unit performance and combat readiness.

III. DISCUSSION

The measurement of training effectiveness, specifically with respect to maintenance training, has been focused on the five aspects of training and the measurement of job performance listed below:

1. Transfer of training
2. Simulation quality
3. Effects of training on individual performance
4. Differential effects of alternative training methods
5. Effects of training and experience on unit performance and operational readiness.

The use of multiple measures of training effectiveness within various studies of F-16 maintenance and performance studies helps to provide a better understanding of the methods of analysis and types of measures that are available for the equipments on which maintenance data have been presented. It is also possible to assess some of the strengths and limitations of the types of performance measures that have been investigated. This includes a consideration of whether they are subjective or objective measures and whether the data are direct or indirect measures of individual or job performance. A summary of the measures that have been reviewed is presented in Table 14.

A. SOURCES OF TRAINING EFFECTIVENESS MEASUREMENT AND EVALUATION

1. Transfer of Training

Two of the studies presented data related to the transfer of training, i.e., how well an individual's training grades or amount of experience (time in service) predict performance in an operational situation. The first study, using a seven-factor, behaviorally anchored rating scale (BARS) (S/PD/JPD) (Wienclaw and Orlansky, 1983), showed that ratings of students in training correlated 0.11 to 0.53 with subsequent ratings as technicians. The second study used a version of the USAF Walk-Through Performance

TABLE 14. SUMMARY OF JOB PERFORMANCE MEASURES OF MAINTENANCE

TYPE OF MEASURE	NAME OF MEASURE	DESCRIPTION	APPLICATION	FINDING	SOURCE
1. S/DPD/JPD	Estimate of Net Job Productivity	Supervisor estimate of job productivity minus the amount of supervisory support required	15 Navy enlisted specialties	Job productivity increased over a four year period.	Quester and Marcus (1985)
2. S/DPD/JPD	Behaviorally Anchored Rating Scale (BARS)	Supervisors rated student and technician performance on 6 scales: Safety, Thoroughness, Use of Technical Data, System Understanding, Mechanical Skills, and Attitude	Air Force F-16 maintenance training with SAMTs and AETs	Job performance increased as time on job increased. Low positive correlation between ratings of students and technicians. Both SAMT and AET groups rated above average as students and as technicians. Rate of on-the-job learning after training appears to be greater for SAMT-trained technicians than for AET-trained technicians.	Wendaw and Orlansky (1983)
3. S/DPD/JPD	Desired Maintenance Result	Weighted profile of job results used to assess maintenance job performance	Air Force F-16 maintenance training with: Hardware Trainer, TFE-14 Gun System Simulator; SAMT, TFE-2, Flight Control; and AETs	Weapon Specialist (AFSC 462X0): Same job performance after TFE-14 or AET training, up to 6 months after FTD. Better job performance after TFE-14 rather than AET training for periods 6 months or longer after FTD. Flight Control/Instrumentation Specialist (AFSC 326X7): Better job performance after AET rather than SAMT training, for periods of 6 months or longer (no data for periods up to 6 months).	Center for Competency Development (1983)
4. S/DPD/JPD	Job Performance Measurement System: Rating Forms	Supervisor, self, and peer rating forms of maintenance job performance	Air Force TF-33 engine maintenance	Supervisor ratings correlated significantly with Total, Hands-on and Interview measures of the WTPT. Peer- and self-completed rating forms showed no consistent relationship to WTPT measures.	Hedge, Ballentine, and Gould (1985)
5. S/DPD/JPI	Troubleshooting Interview	Technician presented with a troubleshooting problem. Responses were rated by SMEs on a 6-point scale.	Air Force F-16 maintenance training with: Hardware Trainer, TFE-14 Gun System Simulator; SAMT, TFE-2, Flight Control; and AETs	Scores improved with on-the-job experience for about a year. Level 5 technicians scored higher and were less variable than level 3 technicians.	Center for Competency Development (1983)

(Continued)

TABLE 14. SUMMARY OF JOB PERFORMANCE MEASURES OF MAINTENANCE (continued)

TYPE OF MEASURE	NAME OF MEASURE	DESCRIPTION	APPLICATION	FINDING	SOURCE
6. S/PI/JPI	"Testimonials"	Personnel from mechanics to Division Officers were interviewed concerning the value of installation level training	Army Installation maintenance training	Users indicated that training is beneficial but could not present supporting data.	McConnell, Buchanan, Johnson, and Murdock (1983)
7. O/PI/JPI	Quality Assurance Personnel Test (QA)	Pass/fail test used for maintenance qualification and certification	Air Force F-4 and F-15 maintenance	F-4 FTD- and OJT-trained personnel had a lower pass rate than "untrained" personnel. F-15 FTD- and OJT-trained personnel had a higher pass rate than "untrained" personnel.	Buchanan, Johnson, and McConnell (1982)
8. O/PI/JPI	Walk-Through Performance Test (WTPT)	Component of the Job Performance Measurement System: 8 Hands-on Tests and 7 Interview Tests	Air Force TF-33 engine maintenance	Hands-on and Interview Tests are significantly correlated with each other ($r = .44$). Training grade, time in unit, remedial instruction, and mechanical aptitude increase multiple R for WTPT total score from .28 to .44.	Hedge, Ballentine, and Gould (1985)
9. O/PI/JPD	Work Unit Code/Trend Analysis	Maintenance man-hours per single procedure performed on a piece of equipment. Score is equal to the ratio of the action time of Work Centers with more trained personnel to the action time of Work Centers with fewer trained personnel	Air Force F-4 and F-15 maintenance	Using proportion of FTD-trained personnel in Work Center as an indicator, Work Centers with more trained personnel complete job actions faster than Work Centers with fewer trained personnel	Buchanan, Johnson, and McConnell (1982)
10. O/PI/JPD	Retest Okay	Percentage of components that were diagnosed and removed as defective which, upon later retesting, prove to function properly	Air Force F-16 wings	Use of measure proved infeasible --RTOK data kept only when system rate exceeds 8% --Many components were incompletely tagged --Many components are used on more than one aircraft --Wing and base are identified but not the Work Center.	McConnell and Johnson (1984)

(Continued)

TABLE 14. SUMMARY OF JOB PERFORMANCE MEASURES OF MAINTENANCE (continued)

TYPE OF MEASURE	NAME OF MEASURE	DESCRIPTION	APPLICATION	FINDING	SOURCE
11. O/P/I/JPD	Productivity	Ratio of the total elapsed time to complete work per job category to the frequency of performing that type of job action	Air Force F-16 maintenance	Productivity increases with the job frequency (experience/workload) in Work Centers with a high percentage of FTD-trained workers. WUC 23Z00 -Productivity increased with training. WUC 23Z00 and 14A00 -Interaction between training and frequency: training plus high frequency yields highest productivity.	Johnson, McConnell, and Murdock (1983)
12. O/P/I/JPD	Average time to complete a work action	Ratio of the work hours per Work Center to the number of work actions per Work Center	Air Force F-16 maintenance	Training significantly improves Work Center performance for 2 out of 3 WUCs. Experience not significant for 3 out of 3 WUCs. Interaction between training and experience: Work Centers with high proportion trained personnel do better with increased workload/experience.	McConnell and Johnson (1984)
13. O/P/I/JPI	A-7 sortie rates	A-7 sortie rates off aircraft carriers for 292 squadron-quarters between 1977 and 1980	Navy A-7 squadrons	Additional junior maintenance personnel seemed to be detrimental to performance. Additional senior maintenance personnel enhanced sortie rate.	Horowitz and Angier (1985)
14. O/P/I/JPI	Ship combat readiness	Navy Casualty Reports (CASREPs) for 91 ships over 3 years	Navy surface ships	Experience level of maintenance personnel is the best predictor of Combat Readiness.	Horowitz and Angier (1985)
15. O/P/I/JPI	Frequency of failure of major equipment	Navy Casualty Reports (CASREPs)	Navy surface ships	Fewer major equipment malfunctions with more experienced maintenance personnel.	Cavalluzzo (1985) Horowitz and Angier (1985)
16. O/P/I/JPI	Training Readiness Score	Training (Combat) Readiness Score (CTRNG) from Navy Unit Status and Identity Report	Navy surface ships	1-day-per-quarter rise in training time is associated with a 2.26 percent rise in the number of ships that are full combat ready upon deployment.	Cavalluzzo (1985)

Test (WTPT) (O/PD/JPI) developed for assessing TF(33) engine maintenance technicians (Hedge, Ballentine, and Gould, 1985). The performance data on the work sample portion of the WTPT correlated significantly with supervisory ratings (S/PD/JPD) ($r = .25$) but not with peer or self ratings. Several training, aptitude, and experience measures were also significantly correlated to the WTPT scores.

Both the BARS and the WTPT data indicate that training has a positive but low relationship to operational performance evaluations. Although using different types of measures, both studies indicated that on-the-job experience after training contributed to improved performance.

2. Influence of Simulation Quality on Effectiveness of Training

The effectiveness of training involving the use of a simulator must be influenced by the quality of the simulator, i.e., the functional similarity between the simulator and actual equipment in areas critical to optimum performance. Therefore, simulation quality must be considered in any program to evaluate training effectiveness where the use of a simulator is one of the performance measurement options under consideration. Given that all simulators are equally effective, it is inevitable that the quality of a particular simulator could have a favorable, neutral, or adverse impact upon the quality of training produced by its use. A simulator that elicits responses that differ from or even conflict with the responses required by the actual equipment should not be expected to be as effective as one that provides a high degree of behavioral fidelity. Two studies have provided information concerning approaches to evaluating simulator fidelity and the effect of fidelity on the relative effectiveness of a suite of simulated maintenance training devices. Jorna and Moraal (1985) demonstrated the importance for training of the correspondence between both the physical and behavioral characteristics of the simulator and the actual equipment. In a maintenance training evaluation, Fitzpatrick and Hritz (1984) compared fidelity ratings by instructors to student confidence ratings and student errors in performing tasks. Student performance errors were lowest on the highest rated trainer and highest on the lowest rated trainer. With only one exception, student confidence ratings and performance errors also mirrored the instructors' relative fidelity ratings on the simulator's operational checks and fault isolation checks.

These two studies show that the judged effectiveness of a simulator is closely related to the correspondence between the simulator task and the actual task. Collection of

data at the task level not only provides a precise and relevant measure of training effectiveness, but also provides a profile of the strengths and weaknesses of the particular simulator being evaluated. Clearly the interpretation of any training effectiveness evaluation of a maintenance-training simulator requires a clear understanding of the device's behavioral fidelity on critical tasks. It would be unwarranted to make any generalizations about the effectiveness of simulator-based training without some consideration of the fidelity of the simulators employed.

3. Effects of Training on Individual Performance

Most of the studies and measures reviewed in Section II were designed to measure the effects of formal training and experience on maintenance performance. A wide variety of methods are used to train military maintenance personnel. Because of the costs in time and effort needed to produce skilled technicians, it is not only reasonable but essential to consider whether the training methods and devices have any real effect on maintenance performance. Eight of the studies presented used maintenance performance measures to determine whether training made any measurable difference in productivity.

Some of the potentially significant long-term effects of formal training programs were presented by Quester and Marcus (1985). Using data from the Enlisted Utilization Survey, which asked supervisors to estimate the net productivity of personnel (work accomplished minus the supervisory time required) (S/PD/JPD), they were able to compare the effects of A-School training and on-the-job training on technician productivity. A-School graduates were more productive than those with only OJT from the end of the first month of operational duty through the end of the four year scope of the study. The OJT personnel started out as less productive and never caught up.

Generally, as a technician gains experience we can expect to see improvements in both quality and speed of work. Two studies tried to collect quality-of-work data and three studies collected speed-of-work data. The quality of work measures were performance on the Quality Assurance Personnel Test (QA) (O/PD/JPI) and the percentage of components removed during maintenance that were later retested okay (O/PI/JPD). The reported attempts to use these measures for performance evaluation proved unsuccessful. It seems that there is a strong "handicapping" variable in operation in the sense that the difficulty level of the test may be adapted to the training or experience of the personnel taking the test.

If this is the case, the "handicapping" needs to be controlled before QA data can be used for evaluating the effects of training or training methods.

McConnell and Johnson (1984) attempted unsuccessfully to collect percent retest-okay data for five F-16 wings. Limitations in the record-keeping practices for component turn-ins made it impossible to trace the turn-ins to the source system, wing, originating Work Center, or individual. The unavailability of data in this effort does not reduce its potential desirability. With better records, percent retest okay should be a good measure of the quality of maintenance performance. Retest okay is known to approach 40 percent in a study that summarized such data but that did not examine the reasons for the observed rates (Orlansky and String, 1981).

Three sequential studies used speed of work as the criterion measure of maintenance productivity. As the precision of the data improved, the quality and quantity of the information to be gained from the data increased. Buchanan et al. (1982) compared the completion times of Work Centers with a higher percentage of FTD-trained personnel with the completion times of Work Centers with lower percentages of FTD-trained personnel (O/PI/JPD). In two F-15 wings and one F-4 wing, there was a small, consistent advantage in favor of the Work Centers that had a higher percentage of FTD-trained personnel.

Using the average elapsed time per worker as a more precise measure of job performance (O/PI/JPD), Johnson et al. (1983) compared the productivity of Work Centers with either more than or less than 60 percent FTD-trained personnel. Work Centers with a higher proportion of FTD-trained personnel were faster than the other Work Centers. In addition, there was an interaction between training and workload such that those Centers with a higher proportion of FTD-trained personnel became more productive under higher workloads. The other Work Centers became less productive under higher workloads. This may be related to an "experience readiness" factor such that the FTD-trained personnel are able to learn from their experience under high workloads and require proportionately less supervisory assistance. In contrast, the less trained personnel may be learning significantly less through experience and increased workloads may overburden the supervisory resources, with a resulting decrease in productivity.

Using average time to complete a work action as the criterion measure (O/PI/JPD), McConnell and Johnson (1984) compared the data for three work unit codes from five

F-16 wings to compare the relative contributions of training and experience on productivity. There was a training by work unit code interaction such that training significantly improved the speed of completing work actions for two of the three work unit codes (23000, jet engine; and 42000, electrical systems) but not for the third (14000, flight control). Experience/workload did not significantly improve performance for any of these work unit codes. However, there appears to be an experience-by-training interaction such that those Work Centers with a higher percentage of FTD-trained personnel performed much faster as a result of increased experience/workload than the Work Centers with a lower proportion of FTD-trained personnel. This again suggests that one product of training is an experience-readiness factor. The FTD-trained personnel seem to benefit more from experience/workload than the untrained, a not unreasonable outcome.

Two studies related the amount of experience of maintenance personnel to the frequency of major equipment problems. Horowitz and Sherman (1977), using Navy Casualty Reports (O/PI/JPI), found that ships with more experienced personnel aboard reported fewer major equipment problems. Horowitz and Angier (1985) found that the fraction of surface combatants with no mission-degrading equipment failures (O/PI/JPI) was related to the ratios of junior and senior maintainers to the number of authorized billets. Adding one senior maintainer (E-5 to E-9) to a ship contributes three times as much to ship readiness as adding a junior one. These data serve to demonstrate that the amount of training/experience among ship personnel has a very real impact on its combat readiness. This is another example of an effect of training that can be deduced from data banks not concerned directly with training.

4. Differential Effects of Training Methods (Simulators vs Actual Equipment)

The comparative differences between using simulators or actual equipment for training has been a major source of concern and controversy in the maintenance training community. There are advocates for the use of simulation or actual equipment despite a scarcity of operational performance data to support either choice. Three of the training effectiveness measurement studies compared the performance of maintenance technicians who were trained either with Simulated Aviation Maintenance Trainers (SAMTs) or with Actual Equipment Trainers (AETs).

The Center for Competency Development (1983) used Troubleshooting Interview (S/PD/JPI) techniques to assess the performance of maintenance specialists at two Air Force bases. The two maintenance specialties used in the study were: AFSC 326X7, Flight Control Specialists; and AFSC 462X0, Weapons Specialists. The relationship between training and performance seems to be specific to the particular Air Force Specialty Codes examined in this study. Troubleshooting scores for Weapon Specialists (462X0) with less than six months operational experience were the same, irrespective of the type of FTD training they had received. The scores of those who had been on the job for over six months were markedly different. The personnel trained with dedicated hardware trainers (HTs) had an average score of 4.5, compared to an average score of 2.8 for the AET personnel. Although the data are cross sectional rather than longitudinal, they suggest that the technicians initially perform about the same, but that the HT technicians progress more rapidly on the job. The use of dedicated aviation maintenance training devices seems to result in a greater "experience readiness."

The troubleshooting data for the Flight Control specialists (326X7) is more limited. The study did not provide any information on the personnel with less than six months of operational experience. For the technician group with over six months of on-the-job experience, the SAMT personnel scored significantly lower than the AET personnel (3.1 vs 3.6). It appears that the simulator training provided for this specialty may not be quite as effective as using AET, although the performance differences are not large. Note that Fitzpatrick and Hritz (1984) reported that the flight control simulator was rated as having lower fidelity than the other F-16 maintenance simulators and it has been consistently related to lower levels of student and technician performance in the reports of Johnson, McConnell, and Murdock (1983) and McConnell and Johnson (1984).

The Troubleshooting Interview data also contained some other performance information of interest. The reported scores exhibited a typical negatively accelerated learning curve with rapid increases in scores during the first year of operational experience after completing FTD training, followed by continuing but less rapid increases for the next six months. The overall magnitude of most of the Troubleshooting Interview Ratings fell into the low to marginally acceptable range. Unfortunately, this result is uninterpretable. It could mean anything from the possibility that training is inadequate to the possibility that the questions were substantially more difficult than the SMEs had estimated. Although

at least some internal evidence to support the position that the questions may have been more difficult than estimated, this cannot really be determined without having a normative distribution of related Troubleshooting Interview questions or an external measure of performance quality.

Additional information on the comparative performance of SAMT- and AET-maintenance personnel is provided in a study which used behaviorally anchored rating scales (BARS) (S/PD/JPD) to evaluate a group of maintainers both as students and as technicians (Wienclaw and Orlansky, 1983). The BARS scores were higher for the AET group both as students and as technicians. This suggests that the AET group may have had some intrinsic advantage that was unrelated to training methods or that the AET is superior to training using SAMTs. Both groups scored higher as technicians than they did as students. This study, and others, indicate that technical skills and performance improve with experience. Of some interest is the fact that the rating gap between the two groups diminished substantially between the times that they were rated as students and as technicians. The SAMT personnel appear to be catching up with the AET personnel. This could be due to a ceiling effect in the rating system, with both groups approaching the ceiling. It could also be due to the SAMT group benefiting more from their on-the-job experience than the AET group--a greater "experience readiness" factor.

The BARS scores for both groups of maintainers were in the highly acceptable category, and the absolute differences between the groups were small. Both methods of training seem to be effective.

The difference in the magnitude of the scores reported by CCD (1983) and by Wienclaw and Orlansky (1983) illustrates the hazards of giving absolute interpretations to ordinal data. The Troubleshooting Interview judged the technicians' responses in relation to a set of ideal solutions. In general, the technicians' performance was judged as poor to marginally acceptable. The BARS scores used supervisor ratings on a set of broad but well-defined categories. The ratings were made within the perceptual set of performance expectations reasonable for students or novice technicians. It is quite possible that the skills and knowledge of an excellent novice may not be much different from those of a marginal journeyman technician. The differences in the absolute performance levels reported by the two studies may be due to characteristics intrinsic to the different measuring

devices; however, the comparisons and trends within the separate studies remain valid and can be generalized.

The third study of interest to the present discussion avoided some of the interpretational problems of the previous studies by using speed of work as the performance measure. McConnell and Johnson (1984) used data obtained from the Maintenance Data Collection and from the Maintenance Management Information and Control System to derive the average time to complete a work action (O/PI/JPD) for three Work Unit Codes (WUCs). The WUCs of interest were: 23000, Jet Engine; 42000, Electrical Systems; and 14000, Flight Control. The data were used to compare Work Center productivity for five F-16 wings.

The composite totals showed that the SAMT-trained personnel performed faster than the AET-trained personnel. It should be noted that the Work Centers with SAMT-trained personnel used in the study had 89 percent FTD-trained personnel, while the Work Centers with AET-trained personnel only had 69 percent FTD-trained personnel; consequently, the observed difference in favor of SAMT training could be exaggerated due to the increased amount as well as the type of training. Analysis of the data by WUCs suggests that there are performance differences due to the training methods and that the differences vary from system to system and between tasks within a system.

The SAMT-trained personnel were faster on two of the WUCs and slower on one WUC than the AET personnel. For WUC 23000, Jet Engines, the SAMT technicians were faster on all four comparison tasks. For WUC 42000, Aircraft Electrical Systems, SAMT technicians were faster on five out of six comparison tasks. For WUC 14000, Flight Control, the SAMT technicians were slower on three out of four comparison tasks. Interestingly, this order of performance essentially replicates the findings by Fitzpatrick and Hritz (1984) in which the engine simulation was rated the highest in fidelity and the avionics/flight control simulation was rated the lowest.

Inspection of the simulator/WUC/task data suggests several trends of interest. First, SAMT personnel consistently perform faster (three out of three WUCs) on the Test-Inspect-Service task. Second, SAMT personnel tend to be slower on the Remove and Replace tasks (two out of three WUCs).

The data generally support the use of simulated aviation maintenance trainers for FTD maintenance training. On the average, the results of their use during FTD training are equal to or better than when actual equipment is used. Personnel trained on the SAMTs judged to be better appear to have an advantage which we may call "experience readiness" which leads to an accelerated improvement in performance when they receive job experience. The data tend to confirm the observation that the quality simulators used in training varies significantly. The personnel trained with the engine and electrical systems simulators tended to do consistently better than their actual equipment-trained counterparts. In contrast, the personnel trained with the avionics/flight control simulator tend to do consistently less well than their actual equipment-trained counterparts. SAMT technicians tend to do consistently well with the Test-Inspect-Service task and less well on the Remove and Replace task.

5. Effects of Training and Experience on Readiness

The effect of training on unit performance and operational readiness provides the final measure of training effectiveness. There is no direct measure in these data to show that a superior method of training, measured by improved performance on the job, contributes more to operational readiness. However, several studies show that personnel training/experience has a significant and meaningful impact on unit performance.

Several reports have used data from navy operations to show the impact of training and experience on the functioning of ships and aircraft. Horowitz and Angier (1985) analyzed A-7 sortie data (O/PI/JPI) and found a positive relation between experience in terms of pay grade and the number of sorties per quarter. In reviewing the Casualty Reports (O/PI/JPI) for 91 ships over a three-year period, they found that experience and training are the most consistent predictors of readiness. Cavalluzzo (1985) using the Training Readiness-Index (CRTRNG) (O/PI/JPI) contained in the Unit Status and Identity Report (UNITREP) found that a one-day-per-quarter increase in training time was associated with a 2.3 percent rise in the number of ships that are reported as full combat ready upon deployment.

At this point we have enough data to show that personnel training and experience do have a demonstrable impact upon accepted measures of unit performance and combat readiness. Some simple quantitative statements regarding the impact of training and experience trade-offs can be made. We now need more refined measures applicable to a

broader range of operational problems. There is also a need for models to relate training requirements and costs to wartime combat readiness. Pellicci (1985) reported the beginnings of a model to be able to specify the training time and costs necessary for an army battalion to achieve combat readiness. This appears to be a step in the right direction. More work needs to be done.

B. INTEGRATION OF MULTIPLE MEASURES OF F-16 MAINTENANCE TRAINING EFFECTIVENESS TO BETTER UNDERSTAND THE MEASURES AND THE SYSTEMS

Many of the studies reviewed used the F-16 maintenance training as the source of the research data. The F-16 represents current aircraft and simulator technology. The Air Force maintains automated maintenance data management systems which can at least track maintenance performance at the Work Center level. The F-16 simulator training systems have been installed incrementally at various Air Force bases. This has, in effect, created a natural field experiment for evaluating the effects of FTD training utilizing either SAMTs or AETs.

A summary of the F-16 maintenance training data is presented in Table 15. Multiple studies which produce the same basic results add credibility to the inferences and conclusions to be drawn from the data. One of the factors to emerge was that the completion of FTD training contributed significantly to productivity for the three Work Unit Codes studied. This finding is contrary to the opinions of a number of supervisors who felt that new personnel basically get all the needed knowledge from job experience prior to completing FTD training (CCD, 1983). Training conducted either with AET or SAMT is capable of producing technicians who are highly rated by their supervisors.

The data summarized in Table 15 provide a basis for making a number of detailed comparisons. With the exception of the Wienclaw and Orlansky (1983) study which evaluated the effects of SAMTs en masse but not individually, only four of the seven types of F-16 SAMTs have been the subject of a published report: TFE-2, Flight Control; TFE-4, Electronics; TFE-11, Engine Diagnostic; and TFE-12, Engine Operating Procedures. None of the studies focuses on the other trainers: TFE-3, Navigation; TFE-10, Engine Start; or TFE-22, Environmental Control. Technician performance was generally faster for those who had been trained with SAMTs than with AETs (three out of four comparisons).

**TABLE 15. MEASURES OF F-16 MAINTENANCE TRAINING EFFECTIVENESS
SUMMARIZED BY SIMULATOR SYSTEM**

SYSTEM/AFSC	MEASURE	FINDING	STUDY
All SAMT-/AET-trained AFSCs	Behaviorally Anchored Rating Scale (BARS) (S/PD/JPD)	<p>Training scores account for 3 to 25% of variance of technician scores</p> <p>Technicians improve over time after training</p> <p>AET personnel rated higher as students and as technicians</p> <p>SAMT personnel improved faster as technicians</p> <p>Differences between groups were small and average ratings were high</p>	Wienclaw and Orlansky (1983)
TFE-4 Electronics SAMT; WUC 42000, Electrical Power Supply; Aircraft Electrical Systems Technician (F-16); AFSC 423X0	<p>Instructor rating of overall fidelity (S)</p> <p>Instructor rating of relative fidelity (S)</p> <p>Student confidence for end-of-course performance (S/PD/JPI)</p> <p>Student end-of-course errors (O/PD-JPI)</p>	<p>Medium fidelity</p> <p>Operational checks better than fault isolation</p> <p>Confidence medium (rank 2nd out of 4)</p> <p>Operational checks better than fault isolation</p> <p>Performance medium (rank 2nd out of 4)</p> <p>Fewer errors on operational checks than on fault isolation</p>	Fitzpatrick and Hritz (1984)
TFE-4 Electronics SAMT; WUC 42000, Electrical Power Supply; Aircraft Electrical Systems Technician AFSC 423X0	Workhours to completion (O/PI/JPD)	<p>Percentage of FTD-trained personnel positively related to productivity</p> <p>Frequency of performance not related to productivity</p> <p>Productivity for SAMT personnel superior to AET personnel on 5 out of 6 tasks: except for Remove and Replace</p>	McConnell and Johnson (1984)

(Continued)

**TABLE 15. MEASURES OF F-16 MAINTENANCE TRAINING EFFECTIVENESS
SUMMARIZED BY SIMULATOR SYSTEM (Continued)**

SYSTEM/AFSC	MEASURE	FINDING	STUDY
TFE-11 Engine Diagnostic; TFE-12 Engine Operating Procedures; WUC 23000 Turbofan Power Plant; Jet Engine Technician (F-16) AFSC 426X4	Instructor rating of overall fidelity (S) Instructor rating of relative fidelity (S) Student confidence for end-of-course performance (S/PD/JPI) Student end of course % errors (O/PD/JPI)	Engine Diagnostic: medium (rank 3rd out of 4) Engine Run: high (rank 1st out of 4) Engine Diagnostic: fault isolation better than operational checks Engine Run: fault isolation better than operational checks Engine Diagnostic: confidence medium (rank 3rd out of 4) Confidence for fault isolation better than for operational checks Engine Run: confidence high (rank 1st out of 4) Confidence for fault isolation higher than for operational checks Engine Diagnostic: % errors medium (rank 3rd out of 4) Lower % errors on fault isolation than on operational checks Engine Run: % errors lowest (rank 1st out of 4) Engine Run: lower % errors on operational checks than on fault isolation	Fitzpatrick and Hritz (1984)
	Elapsed time per worker (O/PI/JPD)	FTD training had a greater effect than experience Effect of training was statistically significant Training by workload interaction: Work Centers with high percentages of FTD personnel increased in productivity with increased workload	Johnson, McConnell, and Murdock (1983)

(Continued)

TABLE 15. MEASURES OF F-16 MAINTENANCE TRAINING EFFECTIVENESS
SUMMARIZED BY SIMULATOR SYSTEM (Continued)

SYSTEM/AFSC	MEASURE	FINDING	STUDY
TFE-11 Engine Diagnostic; TFE-12 Engine Operating Procedures; WUC 23000 Turbofan Power Plant; Jet Engine Technician (F-16) AFSC 426X4 (continued)	Workhours to completion (O/PI/JPD)	Percentage of FTD-trained personnel positively related to productivity Training by workload interaction: Work Centers with high percentages of FTD-trained personnel and higher workloads were the most productive Frequency of performing task was not related to completion time SAMT-trained personnel were faster on all 4 tasks than the AET-trained personnel	McConnell and Johnson (1984)
TFE-14 Hardware Gun System Trainer; WUC 75A00; Weapon System Technician (F-16); AFSC 462X0	Troubleshooting Interview Rating (S/PD/JPI)	< 6 months after FTD training HT and AET equal (2.5 vs 2.5) > 6 months after FTD training HT better than AET (4.5 vs 2.8) Ratings rapidly improved during first year after FTD training Ratings were in the low to marginally acceptable range	Center for Competency Development (1983)
TFE-2 Flight Control/Avionics WUC 14000 Integrated Avionics and Flight Control System Specialist (F-16) AFSC 326X7	Troubleshooting Interview Rating (S/PD/JPI)	> 6 months after FTD AET did better than SAMT (3.6 vs 3.1) Ratings rapidly increased for a year after FTD training Ratings were low to marginally acceptable	Center for Competency Development (1983)

(Continued)

TABLE 15. MEASURES OF F-16 MAINTENANCE TRAINING EFFECTIVENESS
SUMMARIZED BY SIMULATOR SYSTEM (Continued)

SYSTEM/AFSC	MEASURE	FINDING	STUDY
TFE-2 Flight Control/Avionics WUC 14000 Integrated Avionics and Flight Control System Specialist (F-16) AFSC 326X7 (continued)	<p>Instructor rating of overall fidelity (S)</p> <p>Instructor rating of relative fidelity (S)</p> <p>Student confidence for end-of-course performance (S/PD/JPI)</p> <p>Student end of course errors (O/PD/JPI)</p>	<p>Fidelity low (rank 4th out of 4)</p> <p>Operational checks better than fault isolation checks</p> <p>Confidence low (rank 4th out of 4)</p> <p>Performance low (rank 4th out of 4)</p> <p>Lower proportion of errors on operational checks than on fault isolation checks</p>	Fitzpatrick and Hritz (1984)
	Elapsed time per worker (O/PI/JPD)	<p>Training by experience interaction; Work Centers with high percentages of FTD-trained personnel had higher performance with high experience/workload</p> <p>FTD training had a greater effect than experience</p>	Johnson, McConnell, and Murdock (1983)
	Workhours to completion (O/PI/JPD)	<p>Neither the frequency of doing the task nor the percentage of FTD-trained personnel were related to productivity</p> <p>AET personnel were more productive than SAMT personnel on 3 out of 4 tasks</p> <p>The greatest difference in speed was on Remove-and-Replace</p> <p>The SAMT personnel were faster on Test-Inspect-Service</p>	McConnell and Johnson (1984)

SAMT-trained technicians were consistently faster on the Test-Inspect-Service task than the AET-trained technicians. The Remove-and-Replace task was generally performed faster by the AET-trained technicians. On two out of the three WUCs studied, the AET personnel were faster performing the Remove-and-Replace task. The one exception was for WUC 23000, Turbofan Jet Engine. Here it should be noted that the set of trainers for jet engine technicians consists of three SAMTs and a hardware engine trainer.

Some of the F-16 SAMTs were rated highly by the instructors, instilled student confidence, and produced technicians who consistently outperformed their AET counterparts. However, even the least favored of the four most studied SAMTs, the TFE-2, Flight Control, has a number of achievements worth noting: (1) personnel who received their FTD training with this system were significantly more productive than those who had not had FTD training; (2) an interaction between training and experience was observed: personnel who had completed training gained significantly more from experience than those who had not had the FTD training; and (3) the technicians trained with the TFE-2, were faster on the Test-Inspect-Service task than their AET counterparts.

Although the F-16 maintenance system is heavily represented in the recent literature, the evaluation of the effectiveness of the individual training devices is neither systematic nor uniform (see Table 16). The extent of coverage of any single device ranges from zero to six studies and from zero to eight performance measures. While we have learned much about the F-16 maintenance trainers, it would appear that there is much more yet to be learned about the training effectiveness of these and other maintenance training devices.

C. JOB PERFORMANCE MEASURES FOR EVALUATING TRAINING EFFECTIVENESS

The classification scheme used to represent the maintenance training performance measures can be summarized in a 2 x 2 x 2 matrix (observer x subject x task representation). The matrix and the representative measures are presented in tabular form in Table 17. Each measure categorized in the table has value to a potential set of users. Traditional criterion measures used to evaluate personnel selection and training fall within the four Personnel Direct (PD) categories. The four rating scales [Personnel Direct (PD)/Job Performance Direct (JPD)] represent a set of relatively new rating techniques

TABLE 16. SUMMARY OF JOB PERFORMANCE MEASURES AND STUDIES USED TO ASSESS THE TRAINING EFFECTIVENESS OF F-16 MAINTENANCE TRAINING DEVICES

PERFORMANCE MEASURE	STUDY	TRAINING DEVICES										
		TFE-2 FLIGHT CONTROL (SAMT)	TFE-3 NAVIGA- TION (SAMT)	TFE-4 ELEC- TRONICS (SAMT)	TFE-6 SEAT & CANOPY (HT)	TFE-10 ENGINE START (SAMT)	TFE-11 ENGINE DIAG- NOSTICS (SAMT)	TFE-12 ENGINE OPERA- TING (SAMT)	TFE-13 F-100 ENGINE (HT)	TFE-14 GUN (HT)	TFE-15 FUEL (HT)	TFE-22 ENVIRON- MENTAL CONTROL (SAMT)
Behaviorally Anchored Rating Scale (BARS) (S/PD/JPD)	Wiencław & Orlansky (1983)	X	X	X		X	X	X				
Desired Maintenance Result (DMR) (S/PD/JPD)	Center for Competency Development (1983)	X					X			X		
Elapsed Time Per Worker (O/PI/JPD)	Johnson, McConnell, & Murdock (1983)	X					X					
Instructor Ratings of Fidelity (S/PI/JPI)	Fitzpatrick & Hrlitz (1984)	X		X			X	X				
Student Confidence Rating (S/SPD/JPI)	-	X		X			X	X				
Student End-of-Course Errors (O/PI/JPI)	-	X		X			X	X				
Troubleshooting Interview (S/SPD/JPI)	Center for Competency Development (1983)	X								X		
Work Hours to Completion (O/PI/JPD)	McConnell & Johnson (1983)	X		X			X	X				
Number of Measures		8	1	5	0	1	6	5	0	2	0	0
Number of Studies		6	1	3	0	1	4	3	0	2	0	0

designed to provide an accurate measure of how well a technician performs on the job. The data obtained from these measures serve as criterion measures for evaluating personnel selection and training measures. This type of information is useful for personnel management and training, but it does not relate technician performance to unit performance or combat readiness.

The Troubleshooting Interview Rating (Subjective/Personnel Direct/Job Performance Indirect) tries to evaluate individual performance on one of two sample troubleshooting problems. The technique has the advantage of comparing the ratings of performance on a known problem with a textbook solution. The results provide rank order information on how one group of technicians compares with another; however, due to the way the test was developed and used, it is impossible to attach any absolute values to the scores. Since only a small group of problems were administered to a small group of technicians, there is no way to distinguish between problem difficulty and performance quality. For example, uniformly low scores could be the result either of difficult troubleshooting questions, stringent rating standards, or inadequate training.

The Objective/Personnel Direct/Job Performance Indirect category provides direct measures of performance on a representative sample of operational maintenance tasks that a technician is expected to perform. The obtained measures from a carefully constructed device, such as the Walk Through Performance Test, provide a basis for comparing the proficiency of individual technicians but not indicate how well the technicians actually perform on the job. It is still necessary to relate test performance to job performance. This type of performance measurement tends to be expensive to develop and time consuming to use.

The Objective/Personnel Direct/Job Performance Direct category would be an ideal performance measure but is not found in any of the maintenance performance measurement studies reviewed. It would have the advantage of providing an index of the quantity and quality of a technician's work. While it is technically possible to get such measures, it is operationally difficult to do so. Most maintenance is done on a team basis and it is difficult now to trace maintenance actions to a specific person within a Work Center.

Within the Personnel Indirect (PI) category, some measures represent newly available and very useful kinds of information. However, two potential measurement categories can be dispensed with: Subjective/Personnel Indirect/Job Performance Direct is

an empty cell in Table 17 which would include ratings of group performance; and Subjective/Personnel Indirect/Job Performance Indirect, represented by testimonials of training effectiveness, would have little value for assessing the effectiveness of training performance.

TABLE 17. JOB PERFORMANCE MEASUREMENT CHARACTERISTICS

SOURCE OF PERFORMANCE DATA		METHOD OF MEASUREMENT	
PERSONNEL	JOB	SUBJECTIVE	OBJECTIVE
Direct	Direct	Behaviorally Anchored Rating Scale (BARS) - Wienclaw & Orlansky (1983)	"Ideal Category" - no data
"	"	Desired Maintenance Results (DMR) - Center for Competency Development (1983)	
"	"	Job Performance Measurement System: Rating Forms - Hedge, Ballentine, Gould (1985)	
"	"	Supervisory Estimate of Net Job Productivity - Quester & Marcus (1985)	
Direct	Indirect	Troubleshooting Interview - Center for Competency Development (1983)	Quality Assurance Personnel Test (QA) Buchanan, Johnson, & McConnell (1982)
"	"		Job Performance Measurement System: Walk Through Performance Test (WTPT) - Hedge, Ballentine, & Gould (1985)
Indirect	Direct	Group Performance Rating - no data	Elapsed time Per Worker - Johnson, McConnell & Murdock (1983)
"	"		Ratio of Job Completion Times - Buchanan, Johnson & McConnell (1982)
"	"		Retest Okay - McConnell & Johnson (1984)
"	"		Work Hours to Completion - McConnell & Johnson (1984)
Indirect	Indirect	"Testimonials" of the Value of Training - McConnell, Buchanan, Johnson & Murdock (1983)	A-7 Flights Off Carrier - Horowitz & Angier (1985)
"	"		Casualty Reports (CASREPs) - Horowitz & Angier (1985)
"	"		Training Readiness Score (CRTRNG) - Cavalluzzo (1985)

The two objective measurement categories are very useful. Three of the four Objective/Personnel Indirect/Job Performance Direct measures reviewed (Ratio of Job Completion Times, Elapsed Time per Worker, and Hours to Complete Work Action) provided a good basis for evaluating not only the effectiveness of FTD training but also for comparing the comparative strengths and weaknesses of technicians trained with the use of SAMTs or AET. One measure, Retest Okay, failed because of shortcomings in the record-keeping system, but it still remains a good candidate for measuring the quality of work. The Objective/Personnel Indirect/Job Performance Indirect data (Flights Off Carrier, Casualty Reports, and Training Readiness Score) has the immense value of showing the importance of training and experience to unit performance and operational readiness. Since unit performance and operational readiness represent the end products of the maintenance training system, it is important to begin the collection of data and the development of models which show how these end products are affected by personnel and training trade-offs.

The assessment of training effectiveness requires good maintenance job performance data. Some of the measurements reviewed provide an improved capacity for evaluating the effectiveness of training. Clearly, when available, maintenance management data provides a sensitive, unbiased means of evaluating the specific effects of training methods. When objective measures are not reasonably obtainable, subjective measures such as behaviorally anchored rating scales (BARS) and net productivity estimates can provide useful job performance information.

Currently, there is no proven off-the-shelf methodology for collecting job performance data to evaluate maintenance training effectiveness. There are individual efforts which suggest directions for future research. It would be interesting to see the Net Productivity Technique (Quester and Marcus, 1985) and the Behaviorally Anchored Rating Scales (Wienclaw and Orlansky, 1983) used in further investigations. It is important that the assessment of training effectiveness move from the school house to the job site. Certainly the development of job sample tests such as the WTPT is important. However, such job sample tests would be far more useful if it could be demonstrated that they effectively sample the principal factors contributing to maintenance performance effectiveness.

The use of maintenance management data banks as a source of data for evaluating training effectiveness has produced a variety of results. Using work-hours-to-completion data, McConnell and Johnson (1984) produced results which provided some very interesting job-related comparisons of the relative strengths and weaknesses of SAMT and AET training. However, within the same study the attempt to use data bank information for a Retest-Okay analysis proved unsuccessful because the management system did not keep sufficiently detailed records to enable training effectiveness analysis. Given what seem to be both significant strengths and weaknesses, it would be interesting to collect enough of this type of data to see how great an effort is warranted. Despite the promising results thus far, the returns from greater efforts may not justify the amount of effort required.

Of all the literature reviewed, only one performance measure was used for each sample and no two samples used the same measure. It would be useful to see future investigations using multiple measures. This would demonstrate the comparative effectiveness of various measures and whether they sampled the same or different portions of the maintenance performance variance. It is possible that future multiple-measure efforts may sufficiently establish the representativeness of job performance tests that the need for more extensive job performance data from maintenance management data banks will be considerably diminished.

The review of the recent literature on maintenance job performance measures for the assessment of training effectiveness provided the following information on several training issues:

- Training appears to establish both an initial level of proficiency and an improved capacity for more effectively learning from on-the-job experience, termed "experience readiness."
- Training effectiveness studies should assess not only initial job performance but also the rate of change in performance during the first year on the job.
- Objective maintenance job performance data indicated that SAMT-trained technicians were as effective as AET-trained technicians.
- Different training methods were associated with different patterns of strengths and weaknesses. For example,
 - SAMT-trained personnel were consistently faster in performing Test-Inspect-Service tasks
 - AET-trained personnel were consistently faster in performing the Remove-and-Replace tasks.

This review has provided a summary of the research methods and the maintenance performance measures that have been reported in the recent training effectiveness literature. Although the review has been limited to maintenance training, many of the approaches, measurement techniques, and technical insights are applicable to the evaluation of a broad range of training-effectiveness issues. The benefits of using objective data from existing data banks to assess training effectiveness are apparent. Most of the training literature has focused on the early effects of training; increased effort should be directed toward assessing long-term effects of training and experience on individual performance, unit effectiveness, and ultimately, combat readiness.

IV. REFERENCES

- Buchanan, A.O., Johnson, S.C., and McConnell, R.D. (1982) "An Examination of the Impact of a Field Training Detachment (FTD) on Air Force Operational Unit Productivity" (Working Note WN-8202-1). Falls Church, VA 22041: Management Consulting and Research, Inc.
- ✓ Cavalluzzo, L.C. (1985) "Optempo and Training Effectiveness," *Proceedings of the Symposium on Military Value and Cost-Effectiveness of Training* (pp. 197-233). Brussels, Belgium: Defense Research Group Panel on the Defense Applications for Operational Research. (AD B093505)
- Center for Competency Development. (1983) "Training Effectiveness Analysis Study." Coshocton, OH 43812: Author.
- Christensen, J.M. (1948) "Aerial Analysis of Navigator Duties with Special Reference to Equipment Design and Work-Place Layout: II Navigator and Radar Operator Activities During Three Arctic Missions" (Memorandum Report MC REXO-694-15A). Wright-Patterson Air Force Base, OH: USAF, Air Material Command, Engineering Division
- Fitzpatrick, J.A. and Hritz, R.J. (1984) "Acquiring Better Maintenance Trainers: Lessons Learned in the F-16 SAMT Study" AFHRL-83-44. Brooks Air Force Base, Texas 78235: Air Force Human Resources Laboratory.
- Hedge, J.W., Ballentine, R.D., and Gould, R.B. (1985) "Examining the Link Between Training Evaluation and Job Performance Criterion Development (Draft)." *Symposium on Transfer of Training*. Brussels, Belgium: NATO Defense Research Group Panel VIII on the Defense Applications of Human and Bio-Medical Sciences.
- Hedge, J.W., Dickinson, T.L., and Bierstedt, S.A. (1985, August) "The Use of Videotape Technology in Work Sample Testing." Paper presented at the annual meeting of the American Psychological Association. Los Angeles, CA.
- Honeywell. (1984) "Analysis of the Behaviorally Anchored Rating Scale for the F-16 SAMT" (Letter Report). West Covina, CA 91790: Honeywell Training and Control Systems Operation.
- Horowitz, S.A., and Angier, B.N. (1985) "Costs and Benefits of Training and Experience" (pp. 70-96). *Proceedings of the Symposium on Military Value and Cost-Effectiveness of Training*. Brussels, Belgium: Defense Research Group Panel on the Defense Applications of Operational Research. (AD B093505)

Horowitz, S.A., and Sherman, A. (1977) "Crew Characteristics and Ship Condition" (Study 1090). Alexandria, VA: Center for Naval Analyses.

Johnson, S.C., McConnell, R.D., and Murdock, D.H. (1983) "An Examination of the Impact of Installation Level Training Using Simulators on F-16 Unit Maintenance Productivity" (Working Note WN-8202-4). Falls Church, VA 22041: Management Consulting and Research, Inc.

Jorna, P.G.A.M., and Moraal, J. (1985) "The Use of Simulation in Predicting Operational Performance: A New Means Applied to Old Problems (Draft)." Symposium on Transfer of Training. Brussels, Belgium: NATO Defense Research Group Panel VIII on the Defense Applications of Human and Bio-Medical Sciences.

McConnell, R.D., Buchanan, A.O., Johnson, S.C., and Murdock, D.H. (1983) "Skill Training Analysis: The Linkage of Unit Level Skill Training and Unit Productivity" (Technical Report TR-8202-1, Vol. I) and "Skill Training Analysis: An Examination of the Navy Pipeline Management System" (Technical Report TR-8202-1, Vol. II). Falls Church, VA 22041: Management Consulting and Research, Inc.

McConnell, R.D., and Johnson, S.C. (1984) "Skill Training Analysis--Phase II: The Linkage of Unit Level Skill Training and Maintenance Productivity in Air Force F-16 Units" (Technical Report TR-8302-1). Falls Church, VA 22041: Management Consulting and Research, Inc.

✓ McConnell, R.D., and Jones, C.G. (1985) "Skill Training Analysis: An Examination of DOD-Wide Training and Maintenance Systems" (Technical Report TR-8402-1). Falls Church, VA 22041: Management Consulting and Research, Inc.

Mullins, C.J., and Winn, W.R. (Eds.). (1979) "Criterion Development for Job Performance Evaluation: Proceedings from a Symposium" (AFHRL-TR-78-85). Brooks Air Force Base, Texas 78235: Air Force Human Resources Laboratory.

✓ Orlansky, J., and String, J. (1981) "The Performance of Maintenance Technicians on the Job" (IDA Paper P-1597). Arlington, VA 22202: Institute for Defense Analyses. (AD A104347)

✓ Pellicci, J.A. (1985) "Linking Peacetime Training Resources to Wartime Combat Readiness." *Proceedings of the Symposium on Military Value and Cost-Effectiveness of Training* (pp. 333-356). Brussels, Belgium: Defense Research Group Panel on the Defense Applications of Operational Research. (AD B093505)

Quester, A., and Marcus, A. (1985) "An Evaluation of the Effectiveness of Classroom and On-the-Job Training." *Proceedings of the Symposium on Military Value and Cost-Effectiveness of Training* (pp. 197-233). Brussels, Belgium: Defense Research Group Panel on the Defense Applications of Operational Research. (AD B093505)

✓ String, J., and Orlansky, J. (1981) "Evaluating the Effectiveness of Maintenance Training by Using Currently Available Maintenance Data" (IDA Paper P-1574). Arlington, VA 22202: Institute for Defense Analyses. (AD A107843)

Wienclaw, R.A., and Orlansky, J. (1983) "Determining Cost and Training Effectiveness Trade-Offs for Trainer Design: Test of an Experimental Model." *Proceedings of the 5th Interservice/Industrial Training Equipment Conference*, 1, 64-73.

U226943